UNIT 1 COMPUTER ARITHMETIC

Structure **Page Nos.**

1.0 INTRODUCTION

When a calculator or digital computer is used to perform numerical calculations, an unavoidable error, called round-off error must be considered. This error arises because the arithmetic operations performed in a machine involve numbers with only a finite number of digits, with the result that many calculations are performed with approximate representation of actual numbers. The computer has the in–built capability to perform only the basic arithmetic operations of addition, subtraction, multiplication and division. While formulating algorithm all other mathematical operators are reduced to these basic operations even when solving problems involving the operations of calculus. We will discuss the way arithmetic operations are carried out in a computer and some of the peculiarities of computer arithmetics. Finally we dwell on the propagation of errors.

1.1 OBJECTIVES

After going through this unit, you should be able to:

- learn about floating-point representation of numbers;
- learn about non-associativity of arithmetic in computer;
- learn about sources of errors;
- understand the propagation of errors in subsequent calculations;
- understand the effect of loss of significant digits in computation; and
- know when an algorithm is unstable.

1.2 FLOATING POINT ARITHMETIC AND ERRORS

First of all we discuss representation of numbers in floating point format.

1.2.1 Floating Point Representation of Numbers

There are two types of numbers, which are used in calculations:

1. Integers: 1, … −3, −2, −1, 0, 1, 2, 3, …

2. Other Real Numbers, such as numbers with decimal point.

In computers, all the numbers are represented by a (fixed) finite number of digits. Thus, not all integers can be represented in a computer. Only finite number of integers, depending upon the computer system, can be represented. On the other hand, the problem for non-integer real numbers is still more serious, particularly for non-terminating fractions.

Definition 1 (Floating Point Numbers): Scientific calculations are usually carried out in floating point arithmetic in computers.

An n-digit floating-point number in base β (a given natural number), has the form

$$x = \pm (.d_1 d_2 \ldots d_n)_\beta \beta^e,\ 0 \le d_i < \beta,\ m \le e \le M;\ I = 1, 2, \ldots n,\ d_1 \neq 0;$$

where $(.d_1 d_2 \ldots d_n)_\beta$ is a β– fraction called mantissa and its value is given by

$(.d_1 d_2 \ldots d_n)_\beta = d_1 \times \frac{1}{\beta} + d_2 \times \frac{1}{\beta^2} + \quad + d_n \times \frac{1}{\beta^n}$; e is an integer called the exponent.

The exponent e is also limited to range m < e < M, where m and M are integers varying from computer to computer. Usually, m = –M.

In IBM 1130, m = –128 (in binary), –39 (decimal) and M = 127 (in binary), 38 (in decimal).

For most of the computers β = 2 (binary), on some computers β = 16 (hexadecimal) and in pocket calculators β = 10 (decimal).

The precision or length n of floating-point numbers on any computer is usually determined by the word length of the computer.

Representation of real numbers in the computers:

There are two commonly used ways of approximating a given real number x into an n– digits floating point number, i.e. through rounding and chopping. If a number x has the representation in the form $x = (d_1 d_2 \ldots d_{n+1} \ldots)\ \beta^e$, then the floating point number fl(x) in n-digit – mantissa can be obtained in the floating two ways:

Definition 2 (Rounding): fl(x) is chosen as the n–digit floating-point number nearest to x. If the fractional part of $x = d_1 d_2 \ldots d_{n+1}$ requires more than n digits, then if $d_{n+1} < \frac{1}{2}\beta$, then x is represented as $(.d_1\ d_2 \ldots d_n)\ \beta^e$ else, it is written as $(.d_1\ d_2 \ldots d_{n-1}\ (d_n+1)\beta^e$

Example 1: $\text{fl}\left(\frac{2}{3}\right) = .666667 \times 10^0$ in 6 decimal digit floating point representation.

Definition 3 (Chopping): fl(x) is chosen as the floating point number obtained by deleting all the digits except the left-most n digits. Here $d_{n+1}\ldots$ etc. are neglected and $\text{fl}(x) = d_1 d_2 \ldots d_n \beta^e$.

Example 2: If number of digits n = 2, $\text{fl}\left(\frac{2}{3}\right) = (.67) \times 10^0$ rounded

$(.66) \times 10^0$ chopped

$\text{fl}(-83.7) = -(0.84) \times 10^3$ rounded

$-(0.83) \times 10^3$ chopped.

On some computers, this definition of fl(x) is modified in case $|x| \geq \beta^M$ (*overflow*) or $0 < |x| \leq \beta^m$ (*under flow*), where m and M are the bounds on the exponents. Either fl(x) is not defined in this case causing a stop or else fl(x) is represented by a special number which is not subject to the usual rules of arithmetic, when combined with ordinary floating point number.

Definition 4: Let fl(x) be floating point representation of real number x. Then $e_x = |x - fl(x)|$ is called round-off (absolute) error,

$$r_x = \frac{x - fl(x)}{x}$$ is called the relative error.

Theorem: If fl(x) is the n – digit floating point representation in base β of a real number x, then r_x the relative error in x satisfies the following:

(i) $|r_x| < \frac{1}{2}\beta^{1-n}$ if rounding is used.

(ii) $0 \leq |r_x| \leq \beta^{1-n}$ if chopping is used.

For proving (i), you may use the following:

Case 1.

$$d_{n+1} < \frac{1}{2}\beta, \text{ then } fl(x) = \pm(.d_1 d_2 \ldots d_n)\beta^e$$
$$|x - fl(x)| = d_{n+1}, d_{n+2} \ldots \beta^{e-n-1}$$
$$\leq \frac{1}{2}\beta.\beta^{e-n-1} = \frac{1}{2}\beta^{e-n}$$

Case 2.

$$d_{n+1} \geq \frac{1}{2}\beta,$$
$$fl(x) = \pm\{(.d_1 d_2 \ldots d_n)\beta^e + \beta^{e-n}\}$$
$$|x - fl(x)| = .|-d_{n+1}, d_{n+2}.\beta^{e-n-1} + \beta^{e-n}|$$
$$= \beta^{e-n-1}\,|d_{n+1}.\, d_{n+2} - \beta|$$
$$\leq \beta^{e-n-1} \times \frac{1}{2}\beta = \frac{1}{2}\beta^{e-n}$$

1.2.2 Sources of Errors

We list below the types of errors that are encountered while carrying out numerical calculation to solve a problem.

1. Round off errors arise due to floating point representation of initial data in the machine. Subsequent errors in the solution due to this is called propagated errors.

2. Due to finite digit arithmetic operations, the computer generates, in the solution of a problem errors known as generated errors or rounding errors.

3. Sensitivity of the algorithm of the numerical process used for computing f(x): if small changes in the initial data x lead to large errors in the value of f(x) then the algorithm is called *unstable*.

4. Error due to finite representation of an inherently infinite process. For example, consider the use of a finite number of terms in the infinite series expansions of

Sin x, Cos x or f(x) by Maclaurin's or Taylor Series expression. Such errors are called truncation errors.

Generated Error

Error arising due to inexact arithmetic operation is called generated error. Inexact arithmetic operation results due to finite digit arithmetic operations in the machine. If arithmetic operation is done with the (ideal) infinite digit representation then this error would not appear. During an arithmetic operation on two floating point numbers of same length n, we obtain a floating point number of different length m (usually m > n). Computer can not store the resulting number exactly since it can represent numbers a length n. So only n digits are stored. This gives rise to error.

Example 3: Let $a = .75632 \times 10^2$ and $b = .235472 \times 10^{-1}$
$a + b = 75.632 + 0.023$
$= 75.655472$ in accumulator
$a + b = .756555 \times 10$ if 6 decimal digit arithmetic is used.

We denote the corresponding machine operation by superscript * i.e.

$$a +^* b = .756555 \times 10^2 \ (.756555E2)$$

Example 4: Let $a = .23 \times 10^1$ and $b = .30 \times 10^2$

$$\frac{a}{b} = \frac{23}{300} = (0.075666E2)$$

If two decimal digit arithmetic is used then $\frac{a}{b}^* = .76 \times 10^{-1}$ (0.76E – 1)

In general, let w^* be computer operation corresponding to arithmetic operation w on x and y.

Generated error is given by $xwy - xw^*y$. However, computers are designed in such a way that

$xw^*y = fl(xwy)$. So the relative generated error

$$r.g.e. = r_{xwy} = \frac{xwy - xw^* y}{xwy}$$

we observe that in n – digit arithmetic

$$|r.g.e.| < \frac{1}{2} \beta^{1-n}, \text{ if rounding is used.}$$

$$0 \le |\text{r.g.e.}| < \beta^{1-n}, \text{ if chopping is used.}$$

Due to generated error, the associative and the distributive laws of arithmetic are not satisfied in some cases as shown below:

In a computer $3 \times \frac{1}{3}$ would be represented as 0.999999 (in case of six significant digit) but by hand computation it is one. This simple illustration suggested that everything does not go well on computers. More precisely 0.333333 + 0.333333 +0.333333 = 0.999999.

1.2.3 Non-Associativity of Arithmetic

Example 5: Let $a = 0.345 \times 10^0$, $b = 0.245 \times 10^{-3}$ and $c = 0.432 \times 10^{-3}$. Using

3-digit decimal arithmetic with rounding, we have

$b + c = 0.000245 + 0.000432$
$= 0.000677$ (in accumulator)
$= 0.677 \times 10^{-3}$

$a + (b + c) = 0.345 + 0.000677$ (in accumulator)
$= 0.346 \times 10^0$ (in memory) with rounding

$a + b = 0.345 \times 10^0 + 0.245 \times 10^{-3}$
$= 0.345 \times 10^0$ (in memory)

$(a + b) + c = 0.345432$ (in accumulator)
$= 0.345 \times 10^0$ (in memory)

Hence we see that

$$(a + b) + c \neq a + (b + c)$$

Example 6: Let $a = 0.41$, $b = 0.36$ and $c = 0.70$.

Using two decimal digit arithmetic with rounding we have,

$$\frac{(a-b)}{c} = .71 \times 10^{-1}$$

and $\frac{a}{c} - \frac{b}{c} = .59 - .51 = .80 \times 10^{-1}$

while true value of $\frac{(a-b)}{c} = 0.071428 \ldots$.

i.e. $\frac{(a-b)}{c} \neq \frac{a}{c} - \frac{b}{c}$

These above examples show that error is due to finite digit arithmetic.

Definition 5: If x^* is an approximation to x, then we say that x^* approximates x to n significant β digits provided absolute error satisfies

$$\left|x - x^*\right| \leq \frac{1}{2}\beta^{s-n+1},$$

with s the largest integer such that $\beta^s \leq |x|$.

From the above definition, we derive the following:

x^* is said to approximate x correct to n – significant β digits, if

$$\frac{\left|x - x^*\right|}{x} \leq \frac{1}{2}\beta^{1-n}$$

In numerical problems we will use the following modified definition.

Definition 6: x^* is said to approximate x correct to n decimal places (to n places after the decimal)

If $\left|x - x^*\right| \leq \frac{1}{2}10^{-n}$

In n β –digit number, x^* is said to approximate x correct to n places after the dot if $\frac{\left|x - x^*\right|}{x} \leq \beta^{-n}$.

Example7: Let $x^* = .568$ approximate to $x = .5675$

$x - x^* = -.0005$

$$\left|x - x^*\right| = 0.0005 = \frac{1}{2}(.001) = \frac{1}{2} \times 10^{-3}$$

So x^* approximates x correct to 3 decimal place.

Example 8: Let x = 4.5 approximate to x = 4.49998.

$$x - x^* = -.00002$$

$$\frac{\left|x - x^*\right|}{x} = 0.0000044 \leq .000005$$

$$\leq \frac{1}{2}(.00001) = \frac{1}{2}10^{-5} = \frac{1}{2} \times 10^{1-6}$$

Hence, x^* approximates x correct to 6 significant decimal digits.

1.2.4 Propagated Error

In a numerical problem, the true value of numbers may not be used exactly i.e. in place of true values of the numbers, some approximate values like floating point numbers are used initially. The error arising in the problem due to these inexact/approximate values is called propagated error.

Let x^* and y^* be approximations to x and y respectively and w denote arithmetic operation.

The propagated error = $xwy - x^* wy^*$

r.p.e. = relative propagated error

$$= \frac{xwy - x^* wy^*}{xwy}$$

Total Error: Let x^* and y^* be approximations to x and y respectively and let w^* be the machine operation corresponding to the arithmetic operation w. Total relative error

$$r_{xwy} = \frac{xwy - x^* w^* y^*}{xwy}$$

$$= \frac{xwy - x^* wy^*}{xwy} + \frac{x^* wy^* - x^* w^* y^*}{xwy}$$

$$= \frac{xwy - x^* wy^*}{xwy} + \frac{x^* wy^* - x^* w^* y^*}{x^* wy^*}$$

for the first approximation. So total relative error = relative propagated error + relative generated error.

Therefore, $\left|r_{xwy}\right| < 10^{1-n}$ if rounded.
$\left|r_{xwy}\right| < 2.10^{1-n}$ if chopped.
Where $\beta = 10$.

Propagation of error in functional evaluation of a single variable.

Let f(x) be evaluated and x^* be an approximation to x. Then the (absolute) error in evaluation of f(x) is $f(x) - f(x^*)$ and relative error is

$$r_{f(x)} = \frac{f(x) - f(x^*)}{f(x)} \qquad (1)$$

suppose $x = x^* + e_x$, by Taylor's Series, we get $f(x) = f(x^*) + e_x f(x^*) + \ldots$ neglecting higher order term in e_x in the series, we get

$$r_{f(x)} = \frac{e_x f(x^*)}{f(x)} - \frac{e_x}{x} \cong \frac{xf'(x^*)}{f(x)} = r_x \cdot \frac{xf(x^*)}{f(x)}$$

$$|r_{f(x)}| = |r_x| \left| \frac{xf(x^*)}{f(x)} \right|$$

Note: For evaluation of f(x) in denominator of r.h.s. after simplification, f(x) must be replaced by $f(x^*)$ in some cases. So

$$|r_{f(x)}| = |r_x| \left| \frac{xf'(x^*)}{f(x)} \right|$$

The expression $\left| \frac{xf'(x^*)}{f(x)} \right|$ is called condition number of f(x) at x. The larger the condition number, the more ill-conditioned the function is said to be.

Example 9:

1. Let $f(x) = x^{\frac{1}{10}}$ and x approximates x^* correct to n significant decimal digits. Prove that $f(x^*)$ approximates f(x) correct to (n+1) significant decimal digits.

$$r_{f(x)} = r_x \cdot \frac{xf'(x^*)}{f(x)}$$

$$= r_x \cdot \frac{x \cdot \frac{1}{10} x^{*-\frac{9}{10}}}{x^{\frac{1}{10}}}$$

$$= \left(\frac{1}{10}\right) r_x$$

$$|r_{f(x)}| = \left(\frac{1}{10}\right) |r_x| \le \frac{1}{10} \cdot \frac{1}{2} \cdot 10^{1-n} = \frac{1}{2} 10^{1-(n+1)}$$

Therefore, $f(x^*)$ approximates f(x) correct to (n + 1) significant digits.

Example 10: The function $f(x^*) = e^x$ is to be evaluated for any x, $0 \le x \le 50$, correct to at least 6 significant digits. What digit arithmetic should be used to get the required accuracy?

$$|r_{f(x)}| = |r_x| \left| \frac{xf'(x^x)}{f(x)} \right|$$

$$= |r_x| \left| \frac{x.e^{x^*}}{e^x} \right|$$

$$= |r_x| |x|$$

Let n digit arithmetic be used, then

$$|r_x| < \frac{1}{2} 10^{1-n}$$

This is possible, if $|x| \; |r_x| \le \frac{1}{2} 10^{1-6}$

or $50 . \frac{1}{2} 10^{1-n} \le \frac{1}{2} 10^{1-6}$

$$\frac{1}{2}10^{1-n} \le \left(\frac{1}{100}\right)10^{1-6}$$

$$10^{1-n} \le 2.10^{1-8}$$

$$or\ 10^{-n} \le 10^{-8}.2$$

$$-n \le -8 + log_{10}^{2}$$

$$8 - log_{10}^{2} \le n\ or\ 8 - .3 \le n$$

That is $n \geq 8$.

Hence, $n \geq 8$ digit arithmetic must be used.

Propagated Error in a function of two variables.

Let x^* and y^* be approximations to x and y respectively.

For evaluating f(x, y), we actually calculate $f(x^*, y^*)$

$e_{f(x, y)} = f(x, y) - f(x^*, y^*)$

but $f(x, y) = f(x^* + e_x, y^* + e_y)$

$= f(x^*, y^*) + (e_x f_x + e_f f_f)_{(x^*, y^*)}$ – higher order term. Therefore, $e_{f(x, y)} = (e_x f_x + e_f f_f)_{(x^*, y^*)}$. For relative error divide this by f(x, y).

Now we can find the results for propagated error in an addition, multiplication, subtraction and division by using the above results.

(a) **Addition**: $f(x,y) = x + y$

$$e_{x+y} = e_x + e_y$$

$$r_{x+y} = \frac{xe_x}{x(x+y)} + \frac{ye_y}{y(x+y)}$$

$$= r_x\frac{x}{x+y} + r_y\frac{y}{x+y}$$

(b) **Multiplication**: $f(x,y) = xy$

$$e_{x+y} = e_x y + e_y x$$

$$r_{xy} = \frac{e_x}{x} + \frac{e_y}{y}$$

$$= r_x + r_y$$

(c) **Subtraction**: $f(x,y) = x - y$

$$e_{x-y} = e_x y - e_y x$$

$$r_{x-y} = \frac{xe_x}{x(x-y)} - \frac{ye_y}{y(x-y)}$$

$$= r_x\frac{x}{x-y} + r_y\frac{y}{x-y}$$

(d) **Division**: $f(x,y) = \frac{x}{y}$

$$e_{\frac{x}{y}} = e_x.\frac{1}{y} - e_y.\frac{x}{y^2}$$

$$r_{\frac{x}{y}} = \frac{e_x}{x} - \frac{e_y}{y}$$

$$= r_x - r_y$$

1.3 SOME PITFALLS IN COMPUTATIONS

As mentioned earlier, the computer arithmetic is not completely exact. Computer arithmetic sometimes leads to undesirable consequences, which we discuss below:

1.3.1 Loss of Significant Digits

One of the most common (and often avoidable) ways of increasing the importance of an error is known as loss of significant digits.

Loss of significant digits in subtraction of two nearly equal numbers:

The above result of subtraction shows that x and y are nearly equal then the relative error

$$r_{x-y} = r_x \frac{x}{x-y} - r_y \frac{y}{x-y}$$

will become very large and further becomes large if r_x and r_y are of opposite signs.

Suppose we want to calculate the number $z = x - y$ and x^* and y^* are approximations for x and y respectively, good to r digits and assume that x and y do not agree in the most left significant digit, then $z^* = x^* - y^*$ is as good approximation to $x - y$ as x^* and y^* to x and y..

But if x^* and y^* agree at left most digits (one or more) then the left most digits will cancel and there will be loss of significant digits.

The more the digit on left agrees the more loss of significant digits would take place. A similar loss in significant digits occurs when a number is divided by a small number (or multiplied by a very large number).

Remark 1

To avoid this loss of significant digits, in algebraic expressions, we must rationalize and in case of trigonometric functions, Taylor's series must be used.

If no alternative formulation to avoid the loss of significant digits is possible, then carry more significant digits in calculation using floating-using numbers in double precision.

Example 11: Let $x^* = .3454$ and $y^* = .3443$ be approximations to x and y respectively correct to 3 significant digits. Further let $z^* = x^* - y^*$ be the approximation to $x - y$, then show that the relative error in z^* as an approximation to $x - y$ can be as large as 100 times the relative error in x or y.

Solution:

Given, $|r_x|, |r_y|, \leq \frac{1}{2} 10^{1-3}$

$z^* = x^* - y^* = .3454 - .3443$

$= .0011$

$= .11 \times 10^{-2}$

This is correct to one significant digit since last digits 4 in x^* and 3 in y^* are not reliable and second significant digit of i^* is derived from the fourth digits of x^* and y^*.

Max. $|r_z| = \frac{1}{2} 10^{1-1} = \frac{1}{2} = 100. \frac{1}{2} .10^{-2}$

$\geq 100 |r_x|, 100 |r_y|$

Example 12: Let $x = .657562 \times 10^3$ and $y = .657557 \times 10^3$. If we round these numbers then

$x^* = .65756 \times 10^3$ and $y^* = .65756 \times 10^3$. (n = 5)
$x - y = .000005 \times 10^3 = .005$

while $x^* - y^* = 0$, this is due to loss of significant digits.

Now

$$\frac{u}{x-y} = \frac{.253 \times 10^{-2}}{.005} = \frac{253}{500} \neq \frac{1}{2}$$

whereas $\frac{u^*}{x^* - y} = \infty$

Example 13: Solve the quadratic equation $x^2 + 9.9\,x - 1 = 0$ using two decimal digit floating arithmetic with rounding.

Solution:

Solving the quadratic equation, we have

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-9.9 + \sqrt{(9.9)^2 - 4.1.(-1)}}{2}$$

$$= \frac{-9.9 + \sqrt{102}}{2} = \frac{-9.9 + 10}{2} = \frac{.1}{2} = .05$$

while the true solutions are – 10 and 0.1. Now, if we rationalize the expression.

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-4ac}{2a(b + \sqrt{b^2 - 4ac})}$$

$$= \frac{-2c}{b + \sqrt{b^2 - 4ac})} = \frac{2}{9.9 + \sqrt{102}}$$

$$= \frac{2}{9.9 + 10} = \frac{2}{19.9} = \frac{2}{20} \cong .1 \ .(0.1000024)$$

which is one of the true solutions.

1.3.2 Instability of Algorithms

An algorithm is a procedure that describes, an unambiguous manner, a finite sequence of steps to be performed in a specified order. The object of the algorithm generally is to implement a numerical procedure to solve a problem or to find an approximate solution of the problem.

In numerical algorithm errors grow in each step of calculation. Let ε be an initial error and $R_n(\varepsilon)$ represents the growth of an error at the nth step after n subsequence operation due to ε.

If $R_n(\varepsilon) \approx C\,n\,\varepsilon$, where C is a constant independent of n, then the growth of error is called linear. Such linear growth of error is unavoidable and is not serious and the

results are generally accepted when C and ε are small. An algorithm that exhibits linear growth of error is stable.

If $|R_n(\varepsilon)| \approx Ck^n\varepsilon$, $k > 1$, $C > 0$, k and C are independent of n, then growth of error is called exponential. Since the term k^n becomes large for even relatively small values of n. The final result will be completely erroneous in case of exponential growth of error. Such algorithm is called unstable.

Example 14:

Let $y_n = n! \left\{e - \left(1 + \frac{1}{1!} + \frac{1}{2!} + , + \frac{1}{n!}\right)\right\}$ (1)

$$y_n = \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \ldots\ldots \quad (2)$$

$$y_n < \frac{1}{n} + \frac{1}{n^2} + \frac{1}{n^3} + \ldots.$$

$$0 \le y_n < \frac{\frac{1}{n}}{1 - \frac{1}{n}} = \frac{1}{n-1}$$

$y_n \to 0$ as $n \to \infty$

i.e. $\{y_n\}$ is monotonically decreasing sequence which converges to zero. The value of y_9 using (2) is $y_9 = .10991$ correct to 5 significant figures.

Now if we use (1) by writing

$$y_{n+1} = (n+1)! \left\{e - \left(1 + \frac{1}{1!} + \frac{1}{2!} + , + \frac{1}{(n+1)!}\right)\right\}$$

$$i.e.,\ y_{n+1} = (n+1)\, y_n - 1$$

Using (3) and starting with

$y_0 = e - 1 = 1.7183$, we get
$y_1 = .7183$
$y_2 = .4366$
$y_3 = .3098$
$y_4 = .2392$
$y_5 = .1960$
$y_6 = .1760$
$y_7 = .2320$
$y_8 = .8560$
$y_9 = 6.7040$

This value is not correct even to a single significant digit, because algorithm is unstable. This is shown computationally. Now we show it theoretically.

Let y_n^* be computed value by (3), then we have

$y_{n+1} = (n+1)\, y_n - 1$
$y_{n+1}^* = (n+1)\, y_n^* - 1$

$$y_{n+1} - y^*_{n+1} = (n+1)(y_n - y^*_n)$$
$$\text{i.e. } e_{n+1} = (n+1)\, e_n$$
$$e_{n+1} = (n+1)!\, e_o$$
$$|e_{n+1}| > 2^n |e_o| \text{ for } n > 1$$
$$|e_n| > \frac{1}{2}.2^n |e_o|$$

Here k = 2, hence growth of error is exponential and the algorithm is unstable.

Example 15: The integral $E_n = \int_0^1 x^n e^{x-1} dx$ is positive for all $n \geq 0$. But if we integrate by parts, we get $E_n = 1 - nE_n$ $(= x^n e^{x-1} \int_0^1 - \int_0^1 n\, x^{n-1} e^{x-1} dx)$.

Starting from $E_1 = .36787968$ as an approximation to $\frac{1}{e}$ (accurate value of E_1) correct to 7 significant digits, we observe that E_n becomes negative after a finite number of iteration (in 8 digit arithmetic). Explain.

Solution

Let E^*_n be computed value of E_n.

$$E_n - E^*_n = -n(E_{n-1} - E_{n-1})$$
$$e_n = (-1)^n n!\, e_n$$
$$|e_n| \geq \tfrac{1}{2}.2^n |e_o| \text{ hence process is unstable.}$$

Using 4 digit floating point arithmetic and $E_1 = 0.3678 \times 10^0$ we have $E_2 = 0.2650$, $E_3 = 0.2050$, $E_4 = 0.1800$, $E_5 = 0.1000$, $E_6 = 0.4000$. By inspection of the arithmetic, the error in the result is due to rounding error committed in approximating E_2.

Correct values are $E_1 = 0.367879$, $E_2 = 0.264242$. Such an algorithm is known as an unstable algorithm. This algorithm can be made into a stable one by rewriting $E_{n-1} = \frac{1 - E_n}{n}$, n = … 4, 3, 2. This algorithm works backward from large n towards small number. To obtain a starting value one can use the following:

$$E_n = \leq \int_0^1 x^n dx = \frac{1}{n+1}.$$

1.4 SUMMARY

In this unit we have covered the following:

After discussing floating-point representation of numbers we have discussed the arithmetic operations with normalized floating-point numbers. This leads to a discussion on rounding errors. Also we have discussed other sources of errors… like propagated errors loss of significant digits etc. Very brief idea about stability or instability of a numerical algorithm is presented also.

1.5 EXERCISES

E1) Give the floating point representation of the following numbers in 2 decimal digit and 4 decimal digit floating point number using (i) rounding and (ii)

chopping.

(a) 37.21829

(b) 0.022718

(c) 3000527.11059

E2) Show that $a(b - c) \neq ab - ac$
where
$a = .5555 \times 10^1$
$b = .4545 \times 10^1$
$c = .4535 \times 10^1$

E3) How many bits of significance will be lost in the following subtraction?
37.593621 – 37.584216

E4) What is the relative error in the computation of x – y, where x = 0.3721448693 and y = 0.3720214371 with five decimal digit of accuracy?

E5) If x^* approximates x correct to 4 significant decimal figures/digits, then calculate to how many significant decimal figures/digits $e^{x^*/100}$ approximates $e^{x/100}$.

E6) Find a way to calculate

(i) $f(x) = \sqrt{x^2 + 1} - 1$

(ii) $f(x) = x - Sin\, x$

(iii) $f(x) = x - \sqrt{x^2 - \alpha}$

correctly to the number of digits used when it is near zero for (i) and (ii),very much larger than α for (iii)

E7) Evaluate $f(x) = \dfrac{x^3}{x - Sinx}$ when $x = .12 \times 10^{-10}$ using two digit arithmetic.

E8) Let $u = \dfrac{a-b}{c}$ and $v = \dfrac{a}{c} - \dfrac{b}{c}$ when a = .41, b = .36 and c = .70. Using two digit arithmetic show that $|e_v|$ is nearly two times $|e_u|$.

E9) Find the condition number of

(i) $f(x) = \sqrt{x}$

(ii) $f(x) = \dfrac{10}{1 - x^2}$

and comment on its evaluation.

E10) Consider the solution of quadratic equation

$$x^2 + 111.11x + 1.2121 = 0$$

using five-decimal digit floating point chopped arithmetic.

1.6 SOLUTIONS/ANSWERS

E1) (a)

	rounding	**chopping**
	$.37 \times 10^2$	$.37 \times 10^2$
	$.3722 \times 10^2$	$.3721 \times 10^2$

(b) $.23 \times 10^{-1}$ $.22 \times 10^{-1}$
$.2272 \times 10^{-1}$ $.2271 \times 10^{-1}$

(c) $.31 \times 10^{2}$ $.30 \times 10^{2}$
$.3056 \times 10^{2}$ $.3055 \times 10^{2}$

Note: Let x be approximated by
$a_p \ldots a_1 a_0 . a_{-1} a_{-2} \ldots a_{-q}$.
In case $a_{-q-1} > 5$, x is rounded to
$a_p \ldots a_1 a_0 . a_{-1} a_{-2} \ldots (a_{-q} + 1)$

In case $a_{-q-1} = 5$ which is followed by at least one non-zero digit, x is rounded to
$a_p \ldots a_1 a_0 . a_{-1} a_{-2} \ldots a_{-q+1} . (a_{-q} + 1)$

In case $a_{-q-1} = 5$, being the last non-zero digit, x is rounded to
$a_p \ldots a_1 a_0 . a_{-1} a_{-2} \ldots a_{-q}$
if a_{-q} is even or to
$a_p \ldots a_1 a_0 . a_{-1} a_{-2} \ldots a_{-q+1} . (a_{-q} + 1)$
If a_{-q} if odd.

E2) Let
$a = .5555 \times 10^{1}$
$b = .4545 \times 10^{1}$
$c = .4535 \times 10^{1}$

$b - c = .0010 \times 10^{1} = .1000 \times 10^{-1}$

$a(b - c) = (.5555 \times 10^{1}) \times (.1000 \times 10^{-1})$
$= .05555 \times 10^{0}$
$= .5550 \times 10^{-1}$

$ab = (.5555 \times 10^{1})(.4545 \times 10^{1})$
$= (.2524 \times 10^{2})$

$ac = (.5555 \times 10^{1})(.4535 \times 10^{1})$
$= (.2519 \times 10^{2})$

and $ab - ac = .2524 \times 10^{2} - .2519 \times 10^{2}$
$= .0005 \times 10^{2}$
$= .5000 \times 10^{-1}$

Hence $a(b - c) \neq ab - ac$

E3) $37.593621 - 37.584216$
i.e. $(0.37593621)10^{2} - (0.37584216)10^{2}$

Here $x^* = (0.37593621)10^{2}$, $y^* = (0.37584216)10^{2}$
and assume each to be an approximation to x and y, respectively, correct to seven significant digits.

Then, in eight-digit floating-point arithmetic,
$= (0.00009405)10^{2}$
$z^* = x^* - y^* = (0.94050000)10^{-2}$
is the exact difference between x^* and y^*. But as an approximation to $z = x - y$, z^* is good only to three digits, since the fourth significant digit of z^* is derived from the eighth digits of x^* and y^*, and both possibly in error. Here while the error in z^* as an approximation to $z = x - y$ is at most the sum of the errors in x^*

and y^*, the relative error in z^* is possibly 10,000 times the relative error in x^* or y^*. Loss of significant digits is, therefore, dangerous only if we wish to keep the relative error small.

Given $|r_x|, |r_y| < \frac{1}{2} 10^{1-7}$

$z^* = (0.9405) 10^{-2}$

is correct to three significant digits.

Max $|r_z| = \frac{1}{2} 10^{1-3} = 10{,}000 . \frac{1}{2} 10^{-6} \geq 10{,}000 |r_z|, 10{,}000 |r_y|$

E4) With five decimal digit accuracy

$x^* = 0.37214 \times 10^0 \qquad y^* = 0.37202 \times 10^0$

$x^* - y^* = 0.00012$ while $x - y = 0.0001234322$

$$\frac{\left|(x-y)-\left(x^*-y^*\right)\right|}{|x-y|} = \frac{0.0000034322}{0.0001234322} \approx 3 \times 10^{-2}$$

The magnitude of this relative error is quite large when compared with the relative errors of x^* and y^* (which cannot exceed 5×10^{-5} and in this case it is approximately 1.3×10^{-5})

E5) Here $f(x) = e^{x/100}$

$$r_{f(x)} \approx r_x . \frac{x f'(x^*)}{f(x)} \approx r_x \frac{x f'(x^*)}{f(x)} = r_x . e^{x/100} . \frac{1}{100} . \frac{1}{e^{x/100}}$$

i.e.

$$r_{f(x)} \approx \frac{1}{100} |r_x| \leq \frac{1}{100} . \frac{1}{2} 10^{1-4} = \frac{1}{2} 10^{1-6} .$$

Therefore, $e^{x^*/100}$ approximates $e^{x/100}$ correct for 6 significant decimal digits.

E6) (i) Consider the function:

$f(x) = \sqrt{x^2+1} - 1$ whose value may be required for x near 0. Since $\sqrt{x^2+1} \approx 1$ when $x \approx 0$, we see that there is a potential loss of significant digits in the subtraction. If we use five-decimal digit arithmetic and if $x = 10^{-3}$, then $f(x)$ will be computed as 0.

Whereas if we rationalise and write

$$f(x) = \frac{\left(\sqrt{x^2+1}-1\right)\left(\sqrt{x^2+1}+1\right)}{\left(\sqrt{x^2+1}+1\right)} = \frac{x^2}{\sqrt{x^2+1}+1}$$

we get the value as $\frac{1}{2} \times 10^{-6}$

(ii) Consider the function:

$f(x) = x - \text{Sin}\, x$ whose value is required near $x = 0$. The loss of significant digits can be recognised since Sin $x \approx x$ when $x \cong 0$.

To avoid the loss of significance we use the Taylor (Maclaurin) series for $\text{Sin}\, x$

$$\text{Sin}\, x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

Then $f(x) = x - Sin\, x = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} + \ldots$

The series starting with $\frac{x^3}{6}$ is very effective for calculation $f(x)$ when x is small.

(iii) Consider the function:

$$f(x) = x - \sqrt{x^2 - \alpha}$$

as $f(x) = \frac{\left(x - \sqrt{x^2 - \alpha}\right)}{x + \sqrt{x^2 - \alpha}} \left(x + \sqrt{x^2 - \alpha}\right) = \frac{\alpha}{x + \sqrt{x^2 - \alpha}}$

Since when x is very large compared to α, there will be loss of significant digits in subtraction.

E7) $Sinx = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$

$$Sinx = \left(.12 \times 10^{-10}\right) = .12 \times 10^{-10} - .17 \times 10^{-32} + \ldots \approx .12 \times 10^{-10}$$

So $f(x) = \frac{x^3}{x - Sinx} = \infty$

But $f(x) = \frac{x^3}{x - Sinx}$ can be simplified to

$$= \frac{x^3}{\frac{x^3}{3!} - \frac{x^5}{5!} + \ldots} = \frac{1}{\frac{1}{3!} - \frac{x^2}{5!} + \ldots}$$

The value of $\frac{x^3}{x - Sinx}$ for $= .12 \times 10^{-10}$

is $\frac{1}{\frac{1}{3!}} = 6.$

E8) Using two digit arithmetic

$u = \frac{a-b}{c} = .71 \times 10^{-1}$

$v = \frac{a}{c} - \frac{b}{c} = .59 - .51 = .80 \times 10^{-1}$

True value = .071428

$u - fl(u) = |e_u| = .000428$

$v - fl(v) = |e_v| = .0008572$

Thus, $|e_v|$ is nearly two times of $|e_u|$ indicating that u is more accurate than v.

E9) The word condition is used to describe the sensitivity of the function value $f(x)$ to changes in the argument x. The informal formula for Condition of f at x

$$= max\left\{ \frac{f(x)-f(x^*)}{f(x)} \Big/ \left| \frac{x-x^*}{x} \right| : |x-x^*| \text{" small"} \right\}$$

$$\approx \left| \frac{f'(x)x}{f(x)} \right|$$

The larger the condition, the more ill-conditioned the function is said to be.

If $f(x) = \sqrt{x}$, the condition of f is approximately

$$\left| \frac{f'(x)x}{f(x)} \right| = \frac{\left[\frac{1}{2\sqrt{x}} \right] x}{\sqrt{x}} = \frac{1}{2}$$

This indicates that taking square root is a well conditioned process.

But if $f(x) = \frac{10}{1-x^2}$

$$\left| \frac{f^1(x)x}{f(x)} \right| = \frac{20x/(1-x^2)x}{10x/(1-x^2)} = \frac{2x^2}{|1-x^2|}$$

This number can be very large when x is near 1 or –1 signalling that the function is quite ill-conditioned.

E10) Let us calculate

$$x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

$$x_1 = \frac{-111.11+111.09}{2}$$

$$= -0.01000$$

while in fact $x_1 = -0.010910,$ correct to the number of digits shown.

However, if we calculate x_1 as

$$x_1 = \frac{2c}{b+\sqrt{b^2-4ac}}$$

in five-decimal digit arithmetic $x_1 = -0.010910$ which is accurate to five digits.

$$x_1 = \frac{-2 \times 1.2121}{111.11+111.09} = \frac{-2.4242}{222.20}$$

$$= -\frac{24242}{2222000} = -0.0109099 = -.0109099$$

UNIT 2 SOLUTION OF NON-LINEAR EQUATIONS

Structure **Page Nos.**

2.0 INTRODUCTION

In this unit we will discuss one of the most basic problems in numerical analysis. The problem is called a root-finding problem and consists of finding values of the variable x (real) that satisfy the equation $f(x) = 0$, for a given function f. Let f be a real-value function of a real variable. Any real number α for which $f(\alpha) = 0$ is called a root of that equation or a zero of f. We shall confine our discussion to locating only the real roots of f(x), that is, locating non-real complex roots of $f(x) = 0$ will not be discussed. This is one of the oldest numerical approximation problems. The procedures we will discuss range from the classical Newton-Raphson method developed primarily by Isaac Newton over 300 years ago to methods that were established in the recent past.

Myriads of methods are available for locating zeros of functions and in first section we discuss bisection methods and fixed point method. In the second section, Chord Method for finding roots will be discussed. More specifically, we will take up regula-falsi method (or method of false position), Newton-Raphson method, and secant method. In section 3, we will discuss error analysis for iterative methods or convergence analysis of iterative method.

We shall consider the problem of numerical computation of the real roots of a given equation

$$f(x) = 0$$

which may be algebraic or transcendental. It will be assumed that the function f(x) is continuously differentiable a sufficient number of times. Mostly, we shall confine to simple roots and indicate the iteration function for multiple roots in case of Newton Raphson method.

All the methods for numerical solution of equations discussed here will consist of two steps. First step is about the location of the roots, that is, rough approximate value of the roots are obtained as initial approximation to a root. Second step consists of methods, which improve the rough value of each root.

A method for improvement of the value of a root at a second step usually involves a process of successive approximation of iteration. In such a process of successive approximation a sequence $\{X_n\}$ n = 0, 1, 2, … is

generated by the method used starting with the initial approximation x_o of the root α obtained in the first step such that the sequence $\{X_n\}$ converges to α as $n \to \infty$. This x_n is called the nth approximation of nth iterate and it gives a sufficiently accurate value of the root α.

For the first step we need the following theorem:

Theorem 1: If f(x) is continuous in the closed internal [a, b] and f(a) are of opposite signs, then there is at least one real root α of the equation f(x) = 0 such that $a < \alpha < b$.

If further f(x) is differentiable in the open interval (a, b) and either $f'(x) < 0$ or $f'(x) > 0$ in (a, b) then f(x) is strictly monotonic in [a, b] and the root α is unique.

We shall not discuss the case of complex roots, roots of simultaneous equations nor shall we take up cases when all roots are targeted at the same time, in this unit.

2.1 OBJECTIVES

After going through this unit, you should be able to:

- find an approximate real root of the equation f(x) = 0 by various methods;
- know the conditions under which the particular iterative process converges;
- define 'order of convergence' of an iterative method; and
- know how fast an iterative method converges.

2.2 ITERATIVE METHODS FOR LOCATING ROOTS

One of the most frequently occurring problem in scientific work is to find the roots of equations of the form:

$$f(x) = 0$$

In other words, we want to locate zeros of the function f(x). The function f(x) may be a polynomial in x or a transcendental function. Rarely it may be possible to obtain the exact roots of f(x) = 0. In general, we aim to obtain only approximate solutions using some computational techniques. However, it should be borne in mind that the roots can be computed as close to the exact roots as we wish through these methods. We say x^* satisfies f(x) = 0 approximately when $|f(x^*)|$ is small or a point x^* which is close to a solution of f(x) = 0 in some sense like $|x^* - \alpha| < \varepsilon$ where α is a root of f(x) = 0.

To find an initial approximation of the root, we use tabulation method or graphical method which gives an interval containing the root. In this section, we discuss two iterative methods (i) bisection method and (ii) fixed-point method. In a later section we shall discuss about the rate of convergence of these methods.

2.2.1 Bisection Method

Suppose a continuous function f, defined on the interval [a, b], is given, with f(a) and f(b) of opposite signs, i.e. f(a) f(b) < 0, then by Intermediate Value Theorem stated below, there exists a number α on the real line such that $a < \alpha < b$, for which $f(\alpha) = 0$.
Theorem 2 (Intermediate-value Theorem): If the function f is continuous on the closed interval [a, b], and if $f(a) \le y \le f(a)$, then there exists a point c such that $a \le c \le b$ and f(c) = y.

The method calls for a repeated halving of subintervals of [a, b] and, at each step, locating the "half" containing α. To start with, $a_1 = a$ and $b_1 = b$, and let α_1 be the mid point of [a, b], that is $\alpha_1 = \frac{1}{2}(a_1 + b_1)$. If $f(\alpha_1) = 0$, then $\alpha = \alpha_1$. If not, then $f(\alpha_1)$ has the same sign as either $f(a_1)$ or $f(b_1)$. If $f(a_1)\, f(\alpha_1) < 0$, then root lies in (a_1, α_1). Otherwise the root lies in (α_1, b_1). In the first case we set $a_2 = a_1$ and $b_2 = \alpha_1$ and in the later case we set $a_2 = \alpha_1$ and $b_2 = b_1$. Now we reapply the process to the interval (a_2, b_2). Repeat the procedure until the interval width is as small as we desire. At each step, bisection halves the length of the preceding interval. After n steps, the original interval length will be reduced by a factor $\frac{1}{2^n}$ enclosing the root.

Figure 1: Bisection Method

We now mention some stopping procedures that could be applied to terminate the algorithm. Select a tolerance $\varepsilon > 0$ and generate $\alpha_1, \alpha_2, \ldots \alpha_n$ until one of the following conditions is met:

(i) $$|\alpha_n - \alpha_{n-1}| < \varepsilon, \qquad (2.2.1)$$

(ii) $$\frac{|\alpha_n - \alpha_{n-1}|}{|\alpha_n|} < \varepsilon,\ \alpha_n \neq 0,\ or \qquad (2.2.2)$$

(iii) $$|f(\alpha_n)| < \epsilon \qquad (2.2.3)$$

While applying bisection method we repeatedly apply a fixed sequence of steps. Such a method is called an Iteration method.

However, it is pertinent to mention that difficulties can arise using any of these stopping criteria. For example, there exist sequence $\{\alpha_n\}$ with the property that the differences $\alpha_n - \alpha_{n-1}$ converge to zero while the sequence itself diverges. Also it is possible for $f(\alpha_n)$ to be close to zero while α_n differs significantly from α. The criteria given by (2.2.2) is the best stopping criterion to apply since it tests relative error.

Though bisection algorithm is conceptually clear, it has significant drawbacks. It is very slow in converging. But, the method will always converge to a solution and for this reason it is often used to obtain a first approximation for more efficient methods that are going to the discussed.

Theorem 3: Let $f \in C[a, b]$ and suppose $f(a).f(b) < 0$. The bisection procedure generates a sequence $\{\alpha_n\}$ approximating α with the property,

$$|\alpha_n - \alpha| \leq \frac{b-a}{2^n},\ n \geq 1.$$

Now we illustrate the procedure with the help of an example.

Example 1

Find the least positive root of equation.

$$f(x) = x^3 + 4x^2 - 10 = 0$$

Check that f(x) has only one root in the interval in which this least positive root lies. Find α_4 by bisection algorithm.

Solution

Consider the following tables of values.

X	0	1	2
f(x)	–10	–5	14

Take $a_1 = 1$, $b_1 = 2$, since $f(a_1)\, f(b_1) < 0$.

We give the four iterations in the following tabular form.

N	$\mathbf{a_n}$	$\mathbf{b_n}$	$\boldsymbol{\alpha_n}$	$\mathbf{f(\alpha_n)}$
1	1	2	1.5	2.375
2	1	1.5	1.25	– 1.79687
3	1.25	1.5	1.375	0.16211
4	1.25	1.375	1.3125	– 0.84839
5	1.3125	1.375	1.34375	– 0.35098

After four iterations, we have $\alpha_4 = 1.3125$ approximating the root α with an error $|\alpha - \alpha_4| \le |1.375 - 1.325| = .050$ and since $1.3125 < \alpha$.

$$\frac{|\alpha - \alpha_4|}{|\alpha|} < \frac{|b_5 - a_5|}{|a_5|} < \frac{.050}{1.3125} \le \frac{.5}{10} = \frac{1}{2}10^{-1} = \frac{1}{2}10^{1-2}$$

That is, the approximation is correct to at least 2 significant digits.

Remarks 1: Generally the first stage methods for location of the roots of f(x) = 0 are (i) Tabulation method and (ii) Graphical method. The method of tabulation is very crude and labourious and we have used it in the above example to some extent in locating the least positive root of f(x) = 0. In graphical method we plot the graph of the curve y = f(x) on the graph paper and the points where the curve crosses the x-axis gives approximate values of the roots.

2.2.2 Fixed-point Method (or Method of Interation)

This method is also known as Method of Successive Approximations or Method of Iteration. In this method, we write the equation f(x) = 0.

For example $x^3 - x - 1 = 0$ can be written as,

$$x = (1+x)^{\frac{1}{3}}$$

or $$x = \frac{1+x}{x^2}$$

or $$x = \sqrt{\frac{1+x}{x}}$$

Now solving f(x) = 0 is equivalent to solving x = g(x).

Each such g(x) given above is called an iteration function. In fact, these are infinite number of ways in which the original equation f(x) = 0 can be written as x = g(x). Out of all these functions where one is to be selected, will be discussed in the following analysis.

Definition 1: A number ξ is called a fixed point of g(x) if $g(\xi) = \xi$ and g is called the iteration function.

Our problem is now to find out fixed point(s) of g(x). Graphically x = g(x) is equivalent to solving y = x and y = g(x).

Figure 2: Fixed Point Method

Once an iteration function is chosen, to solve x = g(x), we start with some suitable value x_0 close to the root (how to choose this will be explained) and calculate $x_1 = g(x_0)$ (the first approximation), then $x_2 = g(x_1)$ (second approximation) and so on.

In general

$$x_{n+1} = g(x_n),\ n = 0, 1, 2 \ldots$$

The sequence $\{x_n\}$ converges (under some suitable conditions on g) to a number ξ (say). If g is continuous then this gives $\xi = g(\xi)$, that is, ξ is a fixed point of g(x).

Concerning the existence, uniqueness of a fixed point and convergence of the sequence, we state a theorem below:

Theorem 4 (Fixed Point Theorem): Let iteration function g(x) be defined and continuous on a closed interval I = [a, b]. suppose further that g(x) satisfies the following:

(i) $g(x) \in I$ for all $\in I$

(ii) g(x) is differentiable on I = [a, b]
and there exists a non-negative number $k < 1$ such that for all $x \in I$, $|g'(x)| \leq k < 1$.

Then

(a) g(x) has a fixed point ξ,

(b) the fixed point is unique, and

(c) the sequence $\{x_1\}$ generated from the rule $x_{n+1} = g(x_n)$ converges to ξ, the fixed point of g(x), when $x_0 \in [a, b]$

Proof: (a) **Existence:** Suppose g(a) = a or g(b) = b, then there is nothing to be proved. So, suppose g(a)

$\neq a$ and $g(b) \neq b$. Then $g(a) > a$ and $g(b) < b$ since $g(x) \in I$ for all $x \in I$.

Consider $h(x) = g(x) - x$

Then $h(a) = g(a) - a > 0$ and
$h(b) = g(b) - b < 0$

Also $h(x)$ is continuous on I since $g(x)$ is so. Hence by Intermediate Value Theorem, there exists a number ξ, $a < \xi < b$ such that $h(\xi) = 0 \Rightarrow$

$g(\xi) - \xi = 0$, i.e., $g(\xi) = \xi$

Hence $g(x)$ has a fixed point in I.

(b) **Uniqueness**: From (2.2.4)

$h'(x) = g'(x) - 1$, but $|g'(x)| \leq k < 1$

Hence $h'(x) < 0$.

Therefore, $h(x)$ is a decreasing function and it crosses x-axis only once, i.e. $h(x)$ vanishes only once in I.

Therefore $g(x) - x = 0$ only for unique value of x in (a, b). Hence uniqueness.

(c) **Convergence**: Let ξ be the fixed point of $g(x)$. We have
$\xi = g(\xi)$ and $x_{n+1} = g(x_n)$.

Let $e_{n+1} = \xi - x_{n+1} = g(\xi) - g(x_n) = g'(\eta_n)\ (\xi - x_n)$, where η_n lies between x_n and ξ, that is, $e_{n+1} = g'(\eta_n)e_n$.

Thus, we have $|e_{n+1}| \leq k\ |e_n|$. Using this repeatedly

$$|e_n| \leq k^n\ |e_0|$$

$$\lim_{m \to \infty} |e_n| = 0 \text{ since } k < 1,$$

i.e. $\lim_{m \to \infty} |\xi - x_n| = 0 \Rightarrow \{x_n\} \to \xi$

(The sequence $\{x_n\}$ converges to the number ξ)

Hence proved.

Thus, it may be noted that the iterative scheme $x_{n+1} = g(x_n)$ converges under the condition $|g'(x)| < 1$, $x \in [a, b]$.

Example 2

For $x^3 - x - 1 = 0$, find a positive root by the fixed point method. Find minimum number of iterations so that nth approximate x_n is correct to 4 decimal places.

Solution

Write $x = (1 + x)^{\frac{1}{3}} = g(x)$.

The root lies between 1 and 2 since $f(1) = -1$ and $f(2) = 3$.

Also $g(1) - 1 = 2^{\frac{1}{3}} - 1 = +ve$

$g(2) - 2 = 3^{\frac{1}{3}} - 2 = -ve$

$I = [a, b] = [1, 2]$

$g'(x) = \dfrac{1}{3(1+x)^{\frac{2}{3}}}$ is decreasing function and

$$\max_{x \in I} |g'(x)| = g'(1) = \frac{1}{3 \times 2^{\frac{2}{3}}} = k < 1.$$

Since g′(x) = +ve, therefore g(x) is increasing.

$$\max_{x \in I} g(x) = g(2) = 3^{\frac{1}{3}} = 1.442 < 2$$

$$\min_{x \in I} g(x) = g(1) = 2^{\frac{1}{3}} > 1.$$

Hence, g(x) ε I for all x ∈ I.

Therefore, $x_{n+1} = (1 + x_n)^{\frac{1}{3}}$ generates a sequence of numbers which converges to a fixed point of g(x), (starting with $x_o \in I$).

We have k = $\dfrac{1}{3 \times 2^{\frac{2}{3}}} < 1$ and

$|e_n| \leq k^n |e_n|$ and $|e_o| < 1$. Hence for the desired accuracy we have

$$|e_n| \leq \left(\frac{1}{3 \times 2^{\frac{2}{3}}} \right)^n < \frac{1}{2} 10^{-4} \Rightarrow n = 7.$$

Remark 2: In the following figures we observe the importance of g′(x) in the neighbourhood of a fixed point ξ.

Figure 3

In the neighbourhood of ξ, $|g'(x)| > 1$ (the sequences converge in these cases Fig. 3).
In the neighbourhood of ξ, $|g'(x)| < 1$ (the sequences converge in these two cases Fig. 4).

Figure 4

Remark 3: In numerical problems, one may follow the following procedure to find an interval [a, b].

In order to use this method one needs only to see if $|g'(x)|<1$ at a point in the neighbourhood of the root. Therefore, determining an interval I is not necessary.

Choose an interval [a, b] by some trial and check for the following:

(i) a – g(a) and b – g(b) must be of opposite sign (with b – g(b) > 0).
(ii) $|g'(x)| \le k < 1$ for $x \in [a, b]$
(iii) $g'(x)$ is continuous on [a, b].

If above conditions are not satisfied try for a smaller interval and so on.

Example 3

Find the smallest positive root of $e^{-x} - \cos x = 0$ by the fixed point method.

Solution

To locate the smallest positive root, we draw the figures of
$y = e^{-x}$ and $y = \cos x$

Figure 5 **Figure 6**

Figure shows that the desired root lies between 0 and $\frac{\pi}{2}$ i.e. in $(0, \frac{\pi}{2})$.

Now let us try $x = \cos^{-1}(e^{-x}) = g(x)$

$$g'(x) = \frac{1}{\sqrt{e^{2x} - 1}}$$

To make this less than 1, we must choose $e^{2x} - 1 > 1$, that is, $e^{2x} > 2$. This gives $x > \frac{1}{2} \ln 2$ This suggest that we should take the suggested interval $(\frac{1}{2} ln2, \frac{\pi}{2})$, but to take a closed interval, we consider $I = [ln2, \frac{\pi}{2}]$.

Derivative of g(x) implies that g(x) is an increasing function.

$$\max_{x \in I} g(x) = g(ln2) = \cos^{-1}(e^{-ln2})$$

$$= \cos^{-1}\left(\frac{1}{2}\right) = \frac{\pi}{3} = \frac{22}{21} > ln2$$

$$\max_{x \in I} g(x) = g(\frac{\pi}{2}) = \cos^{-1}(e^{-\frac{\pi}{2}}) < \frac{\pi}{2}$$

since $e^{-\frac{\pi}{2}}$ is positive. Hence g(x) $\in$ I for all x $\in$ I.

$$\max_{x\in I} |g'(x)| = \frac{1}{\sqrt{e^{2\ln 2}-1}} = \frac{1}{\sqrt{e^{\ln 4}-1}} = \frac{1}{\sqrt{3}} = k < 1$$

since g'(x) is a decreasing function.

Hence all the sufficient conditions of the Theorem 4 are satisfied.

Now further, suppose that we want to find minimum number of iteration required to get 4 decimal place accuracy. Let n be the minimum number of iterations required for the desired accuracy.

$|e_n| \le k^n |e_o| \le \frac{1}{2} 10^{-4}$

$|e_o| \le \left|\frac{\pi}{2} - \ln 2\right| \le 1.$ Thus the given condition is satisfied if

$\left(\frac{1}{\sqrt{3}}\right)^n \le \frac{1}{2} 10^{-4}$. That is,

$-n \log_{10}\sqrt{3} \le -4 - \log_{10} 2$ i.e.

$n \ge \frac{4.301}{0.238} = 18.07$ i.e., n =19

Example 4: Find the iteration function and interval I = [a, b] which satisfy the conditions of the theorem of fixed point to find the smallest positive root of x = tan x.

Solution:

We rewrite the equation x = tan x as $x = n\pi + \tan^{-1} x, n = 0, \pm 1, \pm 2, ..$

We know that $-\frac{\pi}{2} < \tan^{-1} x < \frac{\pi}{2},$ so for desired root we take n = 1, that is,

We consider x = $\pi + \tan^{-1} x = g(x)$ and

consider I = $\left[\frac{\pi}{2}, \frac{3\pi}{2}\right]$

For x $\in$ $\left[\frac{\pi}{2}, \frac{3\pi}{2}\right]$, we have $-\frac{\pi}{2} < \tan^{-1} x < \frac{\pi}{2}$, hence $\frac{\pi}{2} < g(x) < \frac{3\pi}{2}$

Also $\max_{x\in I} |g'(x)| = \max_{x\in I} \frac{1}{1+x^2} = \frac{1}{1+\frac{\pi^2}{4}} = \frac{4}{4+\pi^2} < 1$

(Since g'(x) is a decreasing function).

Hence for any $x_o \in I = \left[\frac{\pi}{2}, \frac{3\pi}{2}\right]$, the sequence generated by the fixed-point iteration method will converge.

Remark 5: If ξ is a fixed point of g(x) lying in the open interval (c, d) on which $|\varphi'(x)|$ then the sequence $\{x_n\}$ generated with g(x) as iteration function will not converge to ξ, however close x_o to ξ is taken except accidentally. (Consider the root $\xi = 2$ of $f(x) = x^2 - x - 2 = 0$ with $g(x) = x^2 - 2$).

Remark 6: If ξ is a fixed point of g(x) such that $|g'(\xi\xi| = 1,$ then the iteration function with g(x) may or may not converge to ξ. However, if $|g'(\xi\xi| < 1,$ in some deleted neighbourhood of ξ, then it will converge to

ξ, with x_o taken sufficiently close to ξ. If $|g'(\xi\xi)| > 1$, in some deleted neighbourhood of ξ, then sequence will not converge to ξ.

Remark 7: The conditions mentioned in fixed-point theorem are sufficient but not necessary.

Now we discuss one example, which is very simple but conveys the fact that if a function f(x) has more zeros i.e. f(x) = 0 has more than one real root, then we may have to consider different g(x) for different roots.

Example 5: Find the iteration function g(x) and corresponding interval to get the two roots 1 and 2 by fixed point iteration method for the equations

$x^2 - 3x + 2 = 0$

Solution:

(a) For the root x = 1 if we consider $x = \sqrt{3x-2} = g(x)$, then

$g'(x) = \frac{3}{2\sqrt{3x-2}}$ and

$g'(1) = 1$. Hence we choose $g(x) = \frac{x^2+2}{3}$, $I_1 = \left[\frac{1}{2}, \frac{5}{4}\right]$

$g'(x) = \frac{2x}{3} > 0$ for $x \in I_1$. Hence g(x) is increasing. Also $\max_{x \in I_1} |g'(x)| = \frac{5}{6} < 1$.

$\max_{x \in I_1} g(x) = \frac{\frac{25}{16}+2}{3} = \frac{57}{48} < \frac{5}{4}$

$\max_{x \in I_1} g(x) = \frac{\frac{1}{4}+2}{3} = \frac{3}{2} < \frac{1}{2}$

Hence all the conditions of the theorem are satisfied.

(b) Now for the other root 2, consider

If $g(x) = \frac{x^2+2}{3}$, then $g'(2) = \frac{4}{3} > 1$. Hence we choose $g(x) = \sqrt{3x-2}$ with

$I_2 = \left[\frac{3}{2}, \frac{5}{2}\right]$

$g'(x) = \frac{3}{2\sqrt{3x-2}} > 0$ for all $x \in I_2$, so g(x) is increasing.

$\max_{x \in I_2} g(x) = \sqrt{\frac{11}{2}} < \frac{5}{2}$

$\min_{x \in I_2} g(x) = \sqrt{\frac{5}{2}} > \frac{3}{2}$ so g(x) : maps I_2 into itself.

Also $\max_{x \in I_2} |g^1(x)| = \frac{3}{2\sqrt{\frac{9}{2}-2}} = \frac{3}{\sqrt{10}} < 1$ (since g′(x) is a decreasing function).

Hence all the conditions for the fixed point theorem are satisfied.

In the following two examples, we use the corollary to the fixed point theorem (Theorem 4).

Example 6: The equation $f(x) = x^4 - x - 10 = 0$ has a root in the interval [1, 2]. Derive a suitable iteration function $\varphi(x)$ such that the sequence of iterates obtained

from the method $x_{k+1} = \varphi(x_k)$, k = 0, 1, 2, ... converges to the root of f(x)=0. Using this method and the initial approximation $x_0 = 1.8$, iterate thrice.

Solution: Choose $\varphi(x) = (x+10)^{1/4}$, I = [1, 2].

Then $\varphi'(x) = \frac{1}{4}(x+10)^{-\frac{3}{4}} = \frac{1}{4}\cdot\frac{1}{(x+10)^{\frac{3}{4}}}$

$$\max_{x\in I_2} \varphi'(x) = \frac{1}{4}\cdot\frac{1}{(11)^{\frac{3}{4}}}$$

i.e., $\varphi'(x) < 1$ for $x \in [1, 2]$

Given $x_0 = 1.8$

$x_1 = (1.8+10)^{1/4} = 1.8534 = 1.86$

$x_2 = (1.8+10)^{1/4} = 1.8534 = 1.86$

$x_3 = (1.8+10)^{1/4} = 1.8534 = 1.86$

Example 7: The equation $f(x) = x^3 - 5x + 1 = 0$ has a root in the interval [0, 1]. Derive a suitable iteration function $\varphi(x)$, such that the sequence of iterates obtained from the formula $x_{k+1} = \varphi(x_k)$, k = 0, 1, 2, ... converge to the root of f(x)=0. Using this formula and the initial approximation $x_0 = 0.5$, iterate thrice.

Solution: $\varphi(x) = \frac{x^3+1}{5}$ is chosen since $\varphi'(x) = \frac{3x^2}{5}$ and $\max_{0\le x\in I} \varphi'(x) < 1$.

With $x_0 = 0.5$, $x_1 = 0.225 = 0.23$, $x_2 = 0.202$.

What about choosing $\varphi(x) = (5x-1)^{\frac{1}{3}}$?

What is $\max_{0\le x\in I} \varphi'(x)$ in this case?

2.3 CHORD METHODS FOR FINDING ROOTS

In the previous section we have introduced you to two iterative methods for finding the roots of an equation f(x) = 0, namely bisection method and fixed point method. In this section we shall discuss three iterative methods: regula-falsi, Newton-Raphson, and Secant methods. In the next section we shall discuss the efficiency of these methods.

2.3.1 Regula-falsi Method

This method attempts to speed up bisection method retaining its guaranteed convergence. Suppose we want to find a root of the equation f(x) = 0 where f(x) is a continuous function. We start this procedure also by locating two points x_o and x_1 such that $f(x_o)f(x_1) < 0$.

Let us consider the line joining $(x_o, f(x_o))$ and $(x_1, f(x_1))$. This line cuts the x-axis at some point, say x_2. We find $f(x_2)$. If $f(x_2)f(x_0) < 0$, then we replace x_1 by x_2 and draw a straight line connecting $(x_2, f(x_2))$ and $\{x_o, f(x_o)\}$. If $f(x_2)$ and $f(x_o)$ are such that $f(x_2)f(x_0) > 0$, then x_0 is replaced by x_2 and draw a straight line connecting $(x_1, f(x_1))$ and $(x_2, f(x_2))$. Where the straight line crosses x – axis. that print gives x_3. In both the cases, the new interval obtained is smaller then the initial interval. We repeat the above procedure. Ultimately the sequence is guaranteed to converge to the desired root.

Figure 7

The equation of the chord PQ is $y - f(x_0) = \frac{f(x_1) - f(x_o)}{x_1 - x_o}(x - x_o)$

This cuts x-axis at the point x_2 given by

$0 - f(x_o) = \frac{f(x_1) - f(x_o)}{x_1 - x_o}(x_2 - x_o)$

i.e. $x_2 = \frac{x_o f(x_1) - x_1 f(x_o)}{f(x_1) - f(x_o)}$.

In general, $x_{r+1} = \frac{x_{r-1} f(x_r) - x_r f(x_{r-1})}{f(x_r) - f(x_{r-1})}$, r = 1, 2, 3,....

If $f(x_2) = 0$, then x_2 is the required root. If $f(x_2) \neq 0$ and $f(x_o)f(x_2) < 0$, then the next approximation lies in (x_o, x_2). Otherwise it lies in (x_2, x_1). Repeat the process till $|x_{i+1} - x_i| < \varepsilon$.

Example 8: The equation $2x^3 + 5x^2 + 5x + 3 = 0$ has a root in the interval [–2, –1]. Starting with $x_o = -2.0$ and $x_1 = -1.0$ as initial approximations, perform three iteration of the Regula-falsi method.

Solution:

$f(-2) = -16 + 20 - 10 + 3 = -3$

$f(-1) = -2 + 5 - 5 + 3 = 1$, and $f(-2)\, f(-1) < 0$

$x_2 = \frac{x_o f(x_1) - x_1 f(x_o)}{f(x_1) - f(x_o)} = \frac{-2 \times 1 - (-1)(-3)}{1 - (-3)} = \frac{-5}{4}$, i.e.,

$x_2 = -1.25$ (First iteration)

$f(x_2) = \frac{-125}{32} + 5 \times \frac{25}{16} + 5 \times \frac{-5}{4} + 3 = \frac{21}{32}$

The root lies in (x_o, x_2)

$x_3 = -1.384$ or 1.39 (Second iteration) since

$$x_3 = \frac{-2 \times \frac{21}{32} - \left(-\frac{5}{4}\right) \times (-3)}{\frac{21}{32} - (-3)} = \frac{-\frac{42}{32} - \frac{5}{4}}{\frac{21}{32} + 3} = \frac{-\frac{42 - 120}{32}}{\frac{21 + 96}{32}}$$

$$= \frac{-162}{117} = -1.384 \approx -1.39.$$

For next iteration find $f(x_3)$ and proceed in similar fashion.

2.3.2 Newton-Raphson Method

Newton-Raphson method or N–R method in short.

It can be introduced by basing it on the Taylor's expansion as explained below. Let x_o be an initial approximation and assume that x_o is close to the exact root α and

$f'(x_o) \neq 0$. Let $\alpha = x_o + h$ where h is a small quantity in magnitude. Let f(x) satisfy all the conditions of Taylor's theorem. Then

$f(x_o + h) = f(x_o) + h\, f'(x_o) + ...$

The method is derived by assuming that the term involving h^2 is negligible and that $f(x_o) + h\, f'(x_o) = 0$ i.e. $f(x_o) + (\alpha - x_o) f'(x_o) = 0$

$$\text{i.e. } \alpha \approx x_o - \frac{f(x_o)}{f'(x_o)}$$

$$\text{i.e. } x_1 = x_o - \frac{f(x_o)}{f'(x_o)}$$

Geometrically the next approximation, x_1, is the abscissa of the point of intersection of the tangent PT and the x-axis in Figure 8.

The iteration scheme is

$$. \; x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad n = 0, 1, 2, ...$$

Figure 8 : Newton – Raphson Method

N-R method is an extremely powerful technique, but it has a major difficulty – the need to know the value of the derivative of f at each approximation or iteration. derivative evaluation, we discuss a slight variation, known as Secant Method next.

Example 9: Newton-Raphson method is used to find the pth root of a positive real number R. Set up the iteration scheme. Perform three iterations of the method for R=16.0, p =3, starting with the initial approximation 2.0.

Solution: Let us denote pth root of R by x i.e.

$x = R^{1/p}$ or $x^p - R = 0$.

$f'(x) = px^{p-1}$.

Newton-Raphson Iteration scheme is

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},$$

$$= x_k - \frac{x_k^p - R}{p x_k^{p-1}},$$

On simiplification we get $x_{k+1} = \left(1 - \frac{1}{p}\right) x_k + \frac{R}{p\, x_k^{p-1}}, \; k = 0, 1, 2, ...$

For R = 16, p = 3, $x_o = 2$, we get $x_1 = \frac{8}{3} = 2.67$, $x_2 = \frac{91}{36} = 2.53$,

Remark 8: If a root is repeated m times, the N–R method is modified as

$$x_{k+1} = x_k - in\frac{f(x_k)}{f'(x_k)}$$

Example 10 : The quadric equation $x^4 - 4x^2 + 4 = 0$ has a double root. Starting with $x_o = 1.5$, compute two iterations by Newton-Raphson method.

Solution: For m-repeated root of f(x) = 0, the iteration scheme in case of Newton-Raphson method is given by:

$$x_{k+1} = x_k - m\frac{f(x_k)}{f'(x_k)},\ k = 0, 1, 2, \ldots$$

In this case, we have

$$x_{k+1} = x_k - \frac{2\left(x_k^4 - 4x_k^2 + 4\right)}{4x_k^3 - 8x_k},\ (since\ m = 2\ and\ f(x) = x^4 - 4x^2 + 4)$$

With $x_o = 1.5$, we have

$$x_1 = \frac{3}{2} - \frac{2\times\frac{1}{16}}{\frac{3}{2}} = \frac{17}{12} = 1.41$$

Example 11: Perform two iterations of the Newton-Raphson method to find an approximate value of $\frac{1}{15}$ starting with the initial approximation $x_o = 0.02$

Solution: Suppose we want to find the reciprocal of the number N.

Let $f(x) = \frac{1}{x} - N$

Then $f'(x) = \frac{1}{x^2}$ and the iteration scheme is

$$x_{k+1} = x_k - \frac{\frac{1}{x_k} - N}{\frac{-1}{x_k^2}} = 2x_k - Nx_k^2,\ k = 0, 1, 2\ldots$$

In this case we $x_{k+1} = 2x_k - 15\ x_k^2$, $k = 0, 1, 2$. This gives $x_1 = 0.034$, $x_2 = 0.051$, $x_3 = 0.063$, etc.

2.3.3 Secant Method

This method is a modification of the regula-falsi method and retains the use of secants throughout, but dispenses with the bracketing of the root. Given a function f(x) and two given points x_o and x_1,

We compute,

$$x_2 = x_o - \frac{f(x_o)}{\frac{f(x_1) - f(x_o)}{x_1 - x_o}} = \frac{x_o f(x_1) - x_1 f(x_o)}{f(x_1) - f(x_o)}.$$

Figure 9

Above figure illustrates how x_{n+1} is obtained.

Example 12: Apply the Secant method to find a root of the equation

$2x^3 + 3x^2 + 3x + 1 = 0$. Take the initial approximations as $x_o = -0.2$ and $x_1 = -0.4$.

Solutions:

Let $f(x) = 2x^3 + 3x^2 + 3x + 1$

$f(-0.2) = 0.504$

$f(-0.4) = 0.152$

$$x_2 = \frac{x_o f(x_1) - x_1 f(x_o)}{f(x_1) - f(x_o)}, = \frac{.1712}{.352} = 00.48$$

You may now solve the following exercises:

E1) Using the Newton-Raphson method, find the square root of 10 with initial approximation $x_o = 3$.

E2) A fixed point iteration to find a root of $3x^3 + 2x^2 + 3x + 2 = 0$ close to $x_o = -0.5$ is written as $x_{k+1} = -\frac{2 + 3x_k + 2x_k^2}{3x_k^2}$

Does this iteration converge? If so, iterate twice. If not, write a suitable form of the iteration, show that it converges and iterate twice to find the root.

E3) Do three iterations of Secant method to find an approximate root of the equation.
$3x^3 - 4x^2 + 3x - 4 = 0$
Starting with initial approximations $x_o = 0$ and $x_1 = 1$.

E4) Do three iterations of fixed point iteration method to find the smallest positive roots of $x^2 - 3x + 1 = 0$, by choosing a suitable iteration function, that converges. Start with $x_o = 0.5$.

E5) Obtain the smallest positive root of the equation of $x^3 - 5x + 1 = 0$ by using 3 iterations of the bisection method.

E6) Starting with $x_o = 0$, perform two iterations to find an approximate root of the equation $x^3 - 4x + 1 = 0$, using Newton-Raphson method.

E7) Do three iterations of the Secant method to solve the equation
$x^3 + x - 6 = 0$,
starting with $x_o = 1$ and $x_2 = 2$.

E8) Apply bisection method to find an approximation to the positive root of the equation.
$2x - 3 \sin x - 5 = 0$
rounded off to three decimal places.

E9) It is known that the equation $x^3 + 7x^2 + 9 = 0$ has a root between –8 and –7. Use the regula-falsi method to

obtain the root rounded off to 3 decimal places. Stop the iteration when $|x_{i+1} - x_i| < 10^{-4}$

E10) Determine an approximate root of the equation
$\cos x - xe^x = 0$
using Secant method with the two initial approximations as $x_o = 0$ and $x_1 = 1$. Do two iterations

2.4 ITERATIVE METHODS & CONVERGENCE CRITERIA

Let $\{x_n\}$ be a sequence of iterates of a required root α of the equation $f(x) = 0$, generated by a given method.

The error at the nth iteration, denoted by e_n is given by

$$e_n = \alpha - x_n$$

The sequence of iterates $\{x_n\}$ converge to α if and only if $e_n \to 0$ as $n \to \infty$ otherwise the sequence of iterates diverges.

For each method of iteration considered by us, we shall discuss conditions under which the iteration converges.

Let x_0, x_1, x_2, etc be a sequence generated by some iterative method.

2.4.1 Order of Convergence of Iterative Methods

Definition 2: If an iterative method converges, that is, if $\{x_n\}$ converges to the desired root α, and two constants $p \geq 1$ and $C > 0$ exist such that

$$\lim_{n \to \infty} \left| \frac{e_{n+1}}{e_n^p} \right| = C \text{ (C does not depend on n)}$$

then p is called the order of convergence of the method and C is called the asymptotic error constant. An iterative method with higher order of convergence than 1 is expected to converge rapidly. If p = 1, 2, 3, …, then the convergence is called linear, quadratic, cubic… respectively.

(i) For the Fixed Point Iteration method the order of convergence is generally 1, that is, it is of first order (convergence is linear).

(ii) For the Newton-Raphson method, with x_o near the root, the order of convergence is 2, that is, of second order (convergence is quadratic).

(iii) For the Secant Method order of convergence is $1.618 \approx 1.62$ but it is not guaranteed to converge.

The bisection method is guaranteed to converge, but convergence is slow. Regula-falsi method is guaranteed to converge. However, it is slow and order of convergence is 1.

2.4.2 Convergence of a Fixed Point Method

Theorem 5:

If $g'(x)$ is continuous in some neighbourhood of the fixed point ξ of g, then the fixed point method converges linearly provided $g'(\xi) \neq 0$.

Proof: $e_{n+1} = \xi - x_{n+1}$

$= g(\xi) - g(x_n)$

$= g'(\eta_n)(\xi - x_n)$

for some η_n lying between x_n and ξ.

$e_{n+1} = e_n g'(\eta_n)$

Since g'(x) is continuous in a neighbourhood of ξ, we can write

$g'(\eta_n) = g'(\xi) + h_n$ such that $\lim_{n\to\infty} h_n = 0$.

$e_{n+1} = e_n\{g'(\xi) + h_n\}$

On taking $n \to \infty$, we have

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n} = g'(\xi) = C \neq 0$$

Hence p = 1, since $\frac{|e_{n+1}|}{|e_n|} = |g'(\xi)|$

Therefore, fixed point method converges linearly.

Note: Smaller the value of $|g'(\xi)|$, faster would be the convergence.

Theorem 6: If g''(x) is continuous in some neighbourhood of the fixed point ξ of g, then the fixed point method converges quadratically, provided $g'(\xi) = 0$ and $g''(\xi) \neq 0$.

Proof: We have

$e_{n+1} = \xi - x_{n+1}$

$e_{n+1} = g(\xi) - g(x_n)$

$= g(\xi) - g(\xi - e_n)$

By using Taylor's series expansion, we have

$$e_{n+1} = g(\xi) - \{g(\xi) - e_n g'(\xi) + \frac{e_n^2}{2} g''(\eta_n)\}$$

for some η_n lying in the interval of x_n and ξ. That is,

$$e_{n+1} = -\frac{e_n^2}{2}\left[g''(\xi) + h_n\right]$$ since g''(x) is continuous, $h_n \to 0$. as $n \to \infty$.

Thus, $\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^2} = \frac{|g''(\xi)|}{|2|} = C \neq 0$

Here p = 2, hence convergence is quadratic.

Fixed-point iteration is effective when it converges quadratically, as in Newton-Raphson method discussed below.

N–R Method

We define for equation f(x) = 0 an iterative function g(x) as

$g(x) = x - \frac{f(x)}{f'(x)}$, then the method is called Newton's method. We state a theorem without proof which suggests an interval in which if x_o is taken then Newton's method converges. We generate

the sequence $\{x_n\}$ as $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, n =1, 2, 3, …

2.4.3 Convergence of Newton's Method

Theorem 7:

Suppose we are to solve f(x) = 0. If f'(x) $\neq 0$ and f''(x) is continuous on the closed finite interval [a, b] and let the following conditions be satisfied:

(i) f(a) f(b) < 0
(ii) f''(x) is either ≥ 0 or ≤ 0 for all $x \in [a, b]$
(iii) At the end points a, b

$$\frac{|f(a)|}{|f'(a)|} < b - a \text{ and } \frac{|f(b)|}{|f'(b)|} < b - a.$$

Then Newton's method converges to the unique solution ξ of f(x) = 0 in [a, b] for any choice of $x_0 \in [a, b]$.

Theorem 8: Let f(x) be twice continuously differentiable in an open interval containing a simple root ξ of f(x) = 0. Further let f''(x) exists in neighbourhood of. Then the Newton's method converges quadratically.

Proof: $g(x) = x - \frac{f(x)}{f'(x)}$ is continuously differentiable in some open neighbourhood of ξ. On differentiating g(x), we get

$$g'(x) = 1 - \frac{[f'(x)]^2 - f(x)f'(x)}{[f'(x)]^2}$$

$$= \frac{f(x)f''(x)}{[f'(x)]^2}$$

$g''(\xi) = 0$, since $f(\xi) = 0$ ($f'(\xi) \neq 0$ for a simple root ξ)

Also $g''(x) = \frac{-2f(x)[f'(x)][f''(x)]^2 + [f'(x)]^2\{f'(x)f''(x) + f(x)f''(x)\}}{[f'(x)]^4}$

$$= \frac{[f'(x)]^2 f''(x) + f(x)f'(x)f''(x) - 2f(x)[f''(x)]^2}{[f'(x)]^3}$$

$$g''(\xi) = \frac{f''(\xi)}{f'(\xi)} \neq 0$$

By Taylor's formula, we have

$e_{n+1} = \xi - x_{n+1} = g(\xi) - g(x_n)$

$= -g'(\xi)(x_n - \xi) - \frac{1}{2} g''(\eta_n)(x_n - \xi)^2$ for some η_n between ξ and x_n. That is

$$e_{n+1} = g'(\xi)e_n - \frac{1}{2} g''(\eta_n)e^2{}_n$$

since $g'(\xi) = 0$, and $g''(x)$ is continuous, we have

$$e_{n+1} \cong -\frac{1}{2} g''(\xi)\, e_n^2.$$

Hence the Newton's Method converges quadratically if of x_o is chosen sufficiently close to ξ, where ξ is a simple root of ξ.

2.4.4 Rate of Convergence of Secant Method

Suppose f(x) = 0 is to be solved. Consider the curve y = f(x).

Figure 10

Let the chord AB through the points A(x_{n-1}, $f(x_{n-1})$) and B (x_n, $f(x_n)$) be drawn. Suppose this intersects x–axis at C. Denote this value of x by x_{n+1}. That is

$$y - f(x_{n-1}) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} (x - x_{n-1})$$

$y = 0$, $x = x_{n+1}$, we get

$$x_{n+1} = x_{n-1} - \frac{f(x_{n-1})(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}$$

$$= \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

This is known as secant method. The sequence $\{x_n\}$ is generated with starting points x_0, x_1. We get x_2, reject x_0, and use x_1, x_2 to get x_3 and so on.

Let $e_{n+1} = \xi - x_{n+1}$

$$= \xi - \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

writing $e_n = \xi - x_n$, $e_{n-1} = \xi - x_{n-1}$

and $x_n = \xi - e_n$, $x_{n-1} = \xi - e_{n-1}$

we get

$$e_{n+1} = \frac{e_{n-1} f(\xi - e_n) - e_n f(\xi - e_{n-1})}{f(\xi - e_n) - f(\xi - e_{n-1})}$$

$$= e_{n-1} \left\{ f(\xi(- e_n f' (\xi\xi - \frac{e_n^2}{2} f'' (\xi\xi + \text{higher order terms} \right\}$$

$$\frac{- e_n \left\{ f(\xi(- e_{n-1} f' (\xi\xi + \frac{e_n^2}{2} f'' (\xi\xi + \text{higher order terms} \right\}}{f(\xi(- e_n f' (\xi\xi - f(\xi(+ e_{n-1} f' (\xi\xi}$$

since $\because f(\xi) = 0$

$$e_{n+1} \cong \frac{\frac{1}{2}\left(e_n^2 e_{n-1} - e_n e_{n-1}^2\right) f''(\xi\xi}{-\left(e_n - e_{n-1}\right) f'(\xi\xi}$$

$$= -\frac{1}{2}\frac{f''(\xi\xi}{f'(\xi\xi} e_n e_{n-1}, \text{ for sufficiently large n}$$

Let $-\frac{1}{2}\frac{f''(\xi\xi}{f'(\xi\xi} = \alpha$ (fixed constant).

Then $e_{n+1} = \alpha\ e_n e_{n-1}$

Suppose $e_{n+1} = C\ \ e_n^p \Rightarrow e_n = Ce_{n-1}^p$ substituting these in previous results

$$Ce_n^p = \alpha\ e_n \left(\frac{e_n}{C}\right)^{\frac{1}{p}} = \frac{\alpha}{C^{1/p}} e_n^{\,1+\frac{1}{p}}$$

Equating powers of p, on both sides we get

$$P = 1 + \frac{1}{p} \Rightarrow p^2 - p - 1 = 0$$

$$P = \frac{1 \pm \sqrt{5}}{2}, \qquad p = 1.618.$$

Now after comparing the rate of convergence of fixed point method, N–R Method and Secant method, we find that secant is faster than fixed point month and N–R method is faster than secant method. For further qualitative comparison refer the books mentioned.

Apart from the rate of convergence, the amount of computational effort required for iteration and the sensitivity of the method to the starting value and the intermediate values, are two main basis for comparison of various iterative methods discussed here. In the case of Newton's method, if f′ (x) is near zero anytime during the iterative cycle, it may diverge. Furthermore, the amount of computational effort to compute f(x) and f ″ (x) is considerable and time consuming. Whereas the fixed point method is easy to programme.

You may now solve the following exercises.

E11) Let M denote the length of the initial interval $[a_0, b_0]$. Let $(x_0, x_1, x_3 ...)$ represent the successive midpoints generated by the bisection method. Show that $|x_{i+1} - x_i| = \frac{M}{2^{i+2}}$

Also show that the number n of iterations required to generate an approximation to a root to an accuracy ε is given by

$$n > -2 - \frac{\log\left(\varepsilon / M\right)}{\log 2}$$

E12) If $x = \xi$ is a zero of f(x) of order 2, then $f(\xi) = 0$, $f'(\xi) = 0$ and $f''(\xi) \neq 0$. Show that in this case Newton-Raphson's method no longer converges quadratically. Also show that if $f'(\xi) = 0$, $f''(\xi) \neq 0$ and $f'''(x)$ is continuous in the neighbourhood of ξ, the iteration

$$x_{i+1} = x_i - \frac{2f(x_i)}{f(x_i)} = g(x_i)$$

does converge quadratically.

E13) The quadratic equation $x^4 - 4x^2 + 4 = 0$ has a double root at $\sqrt{2}$. Starting with $x_0 = 1.5$, compute three successive approximations to the root by Newton-Raphson method. Do this with $g_1(x) = x - \frac{f(x)}{f'(x)}$ and $g_2(x) = x - \frac{2f(x)}{f'(x)}$ and comment on the order of convergence from your results.

E14) The following are the five successive iterations obtained by the Secant method to find the real positive root of the equation $x^3 - x - 1 = 0$ starting with $x_0 = 1.0$ and $x_1 = 2.0$.

n	2	3	4	5	6	7
x_n	1.166667	1.2531120	1.3372064	1.3238501	1.3247079	1.3247180

Calculate $|e_n|$ and $|e_{n+1}|/|e_n e_{n-1}|$ for n = 2, 3, 4. Also compute the constant directly $\left(f'(\xi)/2f'(\xi)\right)$ assuming the value of ξ correct to eight decimal places as ξ= 1.324718.

E15) If $a_0 = 0$ and $b_0 = 1.0$, how many steps of the bisection method are needed to determine the root with an error of at most 10^{-5}?

2.5 SUMMARY

In this unit we have covered the following points:

The methods for finding an approximate solution of equation in one variable involve two steps:

(i) Find an initial approximation to a root.
(ii) Improve the initial approximation to get more accurate value of the root.

The following iterative methods have been discussed:

(i) Bisection method
(ii) Fixed point iteration method
(iii) Regula-falsi method
(iv) Newton-Raphson method
(v) Secant method

We have introduced the convergence criterion of an iteration process.
We have obtained the order/rate of convergence for the iterative methods discussed.
Finally we have given a comparative performance of these methods.

2.6 SOLUTIONS/ANSWERS

E1) $x = \sqrt{10}$, i.e. $x^2 = 10$. $f(x) = x^2 - 10$

$$x_{n+1} = x_n - \frac{x_n^2 - 10}{2x_n} = \frac{x_n^2 + 10}{2x_n},\ n = 0, 1, 2,$$

$$x_o = 3,\ x_1 = \frac{19}{6} = 3.167,\ x_2 = \frac{(3.167)^2 + 10}{6.334} = 3.162$$

E2) Here $\phi(x) = -\frac{1}{3x^2}\left(2 + 3x + 2x^2\right)$

$\left|\phi'(x)\right| = \left|\frac{1}{3}\frac{(4 + 3x)}{x^3}\right| > 1$ at $x_o = -0.5$

Hence iteration does not converge.

If $\phi(x) = -\frac{1}{3}\left(2 + 2x^2 + 3x^3\right)$, then

$\left|\phi'(x)\right| = \left|-\frac{1}{3}\left(4x + 9x^2\right)\right| < 1$ at $x_o = -0.5$

Hence in this case iteration converges

First iteration $x_1 = -0.708$

Second iteration $x_2 = -0.646$

E3) $f(x) = 3x^3 - 4x^2 + 3x - 4$, $x_o = 0$, $x_1 = 1$.

$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$, n = 1, 2, 3

This gives $x_2 = 2$, $x_3 = 1.167$, $x_4 = 1.255$

E4) Root lies in [0, 1]. We take.

$x = \frac{x^2 + 1}{3} = g(x)$

$g'(x) = \frac{2x}{3} \Rightarrow \left|g'(x)\right| < 1$

Starting with $x_0 = 0.5$, we have

$x_1 = \frac{5}{12} = 0.417$, $x_2 = \frac{169}{432} = 0.391$ and $x_3 = 0.384$

E5) $f(0) > 0$ and $f(1) < 0$. The smallest positive root lies in]0, 1[.

No. of bisection	Bisected value x_i	$f(x_i)$	Improved interval
1	0.5	– 1.375	]0,0.5[
2	0.25	– 0.09375	]0,0.25[
3	0.125	0.37895	]0.125, 0.25[

It is enough to check the sign of $f(x_0)$ – the value need not be calculated.

The approximate value of the desired root is 0.1875.

E6) Here $f(x) = x^3 - 4x + 1$, $x_0 = 0$.

The iteration formula is $x_{i+1} = x_2 - \frac{f(x_x)}{f^1(x_i)}$

i.e. $x_{i+1} = \frac{2x_i^3 - 1}{3x_i^2 - 4}$.

This gives

$x_1 = 0.25, x_1 = 0.254095 \approx 0.2541$

E7) $f(x) = x^3 + x = 6, x_0 = 1, x_1 = 2$

$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$ n = 1, 2, 3, …

This gives $x_2 = 1.5,\ x_3 = 1.609 \approx 1.61,\ x_4 = 1.64.$

E8) Here f(x) = 2x – 3 Sinx – 5

x	0	1	2	2.5	2.8	2.9
f(x)	– 5.0	– 5.51224	– 3.7278	– 1.7954	– 0.4049	0.0822

Thus a positive root lies in the interval [2.8, 2.9].

No. of bisection	Bisected value x_0	$f(x_0)$	Improved Interval
1	2.85	–0.1624	[2.85, 2.9]
2	2.875	–0.0403	[2.875, 2.9]
3	2.8875	–0.02089	[2.875, 2.8875]
4	2.88125		

E9) $$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} = \frac{(-8).9 - (-7)(-55)}{0 + 55} = 7.1406$$

Similarly $x_3 = -7.168174$

The iterated values are presented in tabular form below:

No. of intersections	Interval	Bisected value x_0	The function value $f(x_i)$
1	] –8, –7[	–7.1406	1.862856
2	] –8, –7.1406[	–7.168174	0.358767
3			
4			
5			
6			

Complete the above table. You can find that the difference between the 5th and 6th iterated values is $|7.1748226 - 7.1747855| = 0.0000371$ signaling a stop to the iteration. We conclude that –7.175 is an approximate root rounded to the decimal places.

E10) Here $f(x) = \cos x - xe^2,\ x_0 = 0$ and $x_1 = 1$

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} = 0.3146653378$$

$$x_3 = \frac{x_1 f(x_2) - x_2 f(x_1)}{f(x_2) - f(x_1)} = 0.4467281466$$

E11) Starting with bisection method with initial interval $[a_0, b_0]$ (recall that in each step the interval width is reduced by $\frac{1}{2}$ we have

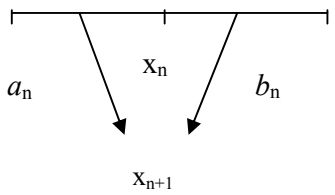
$$b_1 - a_1 = \frac{b_0 - a_0}{2} = \frac{M}{2}$$

$$b_2 - a_2 = \frac{b_1 - a_1}{2} = \frac{b_0 - a_0}{2}$$

and finally $b_n - a_n = \frac{b_0 - a_0}{2^n}$

Let $x_n = \frac{a_n - b_n}{2}$.

$\frac{a_n + b_n}{2}$



Then $x_{n+1} - x_n = \frac{a_{n+1} + b_{n+1}}{2} - \frac{a_n + b_n}{2}$

We have either $a_{n+1} = \frac{a_n + b_n}{2}$ and $b_{n+1} = b_n$

or $a_{n+1} = a_n$ and $b_{n+1} = \frac{a_n + b_n}{2}$.

In either case

$$|x_{n+1} - x_n| = \frac{b_n - a_n}{2^2} = \frac{b_0 - a_0}{2^{n+2}}$$

We want $|x_{n+1} - x_n| = \frac{M}{2^{n+2}} < \varepsilon$.

This is satisfied if

$$log\left(\frac{M}{2^{n+2}}\right) < log\varepsilon$$

log M – (n + 2) log 2 < log ε
nlog2 > – 2log2 + log M – log ε

$$n > -2 - \frac{log\left(\varepsilon / M\right)}{log2}$$

E12) In case we have for a simple root ξ

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = g(x_n)$$

$g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}$. Thus

$g'(\xi) = 0$ since f(ξ) = 0 and $g''(\xi) = \frac{f''(\xi)}{f'(\xi)} (f'(\xi) \neq 0)$

But given that f(ξ) = 0 = f'(ξ) and f''(ξ) $\neq$ 0.

In this case

$$\lim_{x \to \xi} \frac{f(x)f''(x)}{[f'(x)]^2} = \lim_{x \to \xi} \frac{f(x)}{[f'(x)]^2} \lim_{x \to \xi} f''(x)$$

$$= \lim_{x \to \xi} \frac{f'(x)}{2f(x)f''(x)} \lim_{x \to \xi} f''(x) = \frac{1}{2}$$ (by L'Hospital Rule)

That is $g'(\xi) = \frac{1}{2} \neq 0$

Hence it does not converge quadratically.

In case

$$x_{n+1} = x_n - \frac{2f(x_n)}{f'(x_n)} \text{ where } g(x) = x - \frac{2f(x)}{f'(x)}$$

$$g'(x) = \frac{2f(x)f''(x) - [f'(x)]^2}{[f'(x)]^2} \text{ and } g'(\xi\xi = 0 .$$

Since

$$g'(x) = \frac{2f(x)f''(x)}{[f'(x)]^2} - 1 \text{ and}$$

$$\lim_{x\to\xi} g'(x) = \lim_{x\to\xi} \frac{2f(x)f''(x)}{[f'(x)]^2} - 1$$

$$= \lim_{x\to\xi} \frac{2f(x)}{[f'(x)]^2} \cdot \lim_{x\to\xi} f''(x) - 1$$

$$= 2\times\frac{1}{2} - 1 = 0.$$

E13) $f(1.5) = 5.0625 - 9 + 4 = .0625$
$f'(1.5) = 13.5 - 12 = 1.5$

With $g_1(x)$

$x_1 = 1.5 - \frac{.0625}{1.5} = 1.5 - .04 = 1.46$
$f(1.46) = 4.543 - 8.52 + 4 = 0.02$
$f'(1.46) = 12.45 - 11.68 = 0.77$
$x_2 = 1.46 - \frac{0.02}{0.77} = 1.44$

With $g_2(x)$

$x_1 = 1.5 - \frac{2\times.0625}{1.5} = 1.5 - 0.08 = 1.42$
$f(1.42) = 4.065 - 8.065 + 4 = 0$
$x_2 = 1.42$
Actual root = 1.4142. Hence convergence is faster with $g_2(x)$ with two decimal digit arithmetic.

E14) We have the following results in tabular form:

n	1	2	3	4	5
e_n	0.1580513	0.0716060	0.012884	0.0008679	0.0000101
$\lvert e_{n+1}\rvert / \lvert e_n e_{n-1}\rvert$		1.1034669	0.9705400	0.9318475	

Also $f''(\xi\xi/2f'(\xi) = 0.93188$ $(\xi = 1.3247180)$ $\lvert e_5\rvert/\lvert e_4 e_3\rvert = 0.9318475.$ Hence agreement is good for n = 4.

E15) Here M = $b_0 - a_0 = 1 - 0 = 1$

$$n > -2 - \frac{\log(\varepsilon/M)}{\log 2}$$

Here $\varepsilon = 10^{-5}$

So $n > -2 - \frac{log10^{-5}}{log2}$

$= -2 + \frac{5log10}{log2}$

$= -2 + \frac{5}{log2}$

$= -2 + 5 \times 3.322$

$= -2 + 16.61$

$= 14.61$

n $\geq$ 15

UNIT 1 SOLUTION OF LINEAR ALGEBRAIC EQUATIONS

Structure **Page Nos.**

1.0 INTRODUCTION

In Block 1, we have discussed various numerical methods for finding the approximate roots of an equation $f(x) = 0$. Another important problem of applied mathematics is to find the (approximate) solution of systems of linear equations. Such systems of linear equations arise in a large number of areas, both directly in the modelling physical situations and indirectly in the numerical solution of other mathematical models. Linear algebraic systems also appear in the optimization theory, least square fitting of data, numerical solution of boundary value problems of ODE's and PDE's etc.

In this unit we will consider two techniques for solving systems of linear algebraic equations – Direct method and Iterative method.

These methods are specially suited for computers. Direct methods are those that, in the absence of round-off or other errors, yield the exact solution in a finite number of elementary arithmetic operations. In practice, because a computer works with a finite word length, direct methods do not yield exact solutions.

Indeed, errors arising from round-off, instability, and loss of significance may lead to extremely poor or even useless results. The fundamental method used for direct solution is Gauss elimination.

Iterative methods are those which start with an initial approximations and which, by applying a suitably chosen algorithm, lead to successively better approximations. By this method, even if the process converges, we can only hope to obtain an approximate solution. The important advantages of iterative methods are the simplicity and uniformity of the operations to be performed and well suited for computers and their relative insensivity to the growth of round-off errors.

So far, you know about the well-known Cramer's rule for solving such a system of equations. The Cramer's rule, although the simplest and the most direct method, remains a theoretical rule since it is a thoroughly inefficient numerical method where even for a system of ten equations, the total number of arithmetical operations required in the process is astronomically high and will take a huge chunk of computer time.

1.1 OBJECTIVES

After going through this unit, you should be able to:

- obtain the solution of system of linear algebraic equations by direct methods such as Cramer's rule, and Gauss elimination method;
- use the pivoting technique while transforming the coefficient matrix to upper triangular matrix;
- obtain the solution of system of linear equations, $A\mathbf{x} = \mathbf{b}$ when the matrix A is large or spare, by using one of the iterative methods – Jacobi or the Gauss-Seidel method;
- predict whether the iterative methods converge or not; and
- state the difference between the direct and iterative methods.

1.2 PRELIMINARIES

Let us consider a system of n linear algebraic equations in n unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots &\quad + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots &\quad + a_{2n}x_n = b_2 \\ a_{n1}x_1 + a_{n2}x_2 + \ldots &\quad + a_{nn}x_n = b_n \end{aligned} \qquad (1.2.1)$$

Where the coefficients a_{ij} and the constants b_i are real and known. This system of equations in matrix form may be written as

$$A\mathbf{x} = \mathbf{b} \quad \text{where } A = (a_{ij})_{n \times n} \qquad (1.2.2)$$
$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \text{ and } \mathbf{b} = (b_1, b_2, \ldots, b_n)^T.$$

A is called the coefficient matrix.

We are interested in finding the values x_i, i = 1, 2… n if they exist, satisfying Equation (3.3.2).

We now give the following

Definition 1: A matrix in which all the off-diagonal elements are zero, i.e. $a_{ij} = 0$ for i $\neq$ j is called a diagonal matrix; e.g., $A = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$ is a 3 × 3 diagonal matrix.

A square matrix is said to be upper – triangular if
$a_{ij} = 0$ for i > j, e.g.,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

Definition 2: A system of linear equations (3.3.2) is said to be consistent thus exists a solution. The system is said to be inconsistent if no solution exists. The system of equations (3.3.2) is said to be homogeneous if vector $\underline{b} = \underline{0}$, that is, all $b_i = 0$, otherwise the system is called non-homogeneous.

We state the following useful result on the solvability of linear systems.

Theorem 1: A non-homogeneous system of n linear equations in n unknown has a unique solution if and only if the coefficient matrix A is non singular (det A ≠ 0) and the solution can be expressed as $\mathbf{x} = A^{-1}\mathbf{b}$.

1.3 DIRECT METHODS

In schools, generally Cramer's rule/method is taught to solve system of simultaneous equations, based on the evaluation of determinants. This is a direct method. When n is small (say, 3 or 4), this rule is satisfactory. However, the number of multiplication operations needed increases very rapidly as the number of equations increases as shown below:

Number of equations	**Number of multiplication operations**
2	8
3	51
4	364
5	2885
.	
.	
.	
10	359251210

Hence a different approach is needed to solve such a system of equations on a computer. Thus, Cramer's rule, although the simplest and the most direct method, remains a theoretical rule and we have to look for other efficient direct methods. We are going to discuss one such direct method – Gauss' elimination method next after stating Cramer's Rule for the sake of completeness.

1.3.1 Cramer's Rule

In the system of equation (3.3.2), let $\Delta = \det(A)$ and $\mathbf{b} \neq 0$. Then the solutions of the system is obtained as $x_i = \Delta_i / \Delta$, i = 1, 2, … , n

where Δ_i is the determinant of the matrix obtained from A by replacing the i^{th} column of Δ by vector **b**.

1.3.2 Gauss Elimination Method

In Gauss's elimination method, one usually finds successively a finite number of linear systems equivalent to the given one such that the final system is so simple that its solution may be readily computed. In this method, the matrix A is reduced to the form U (upper triangular matrix) by using the elementary row operations like

(i) interchanging any two rows
(ii) multiplying (or dividing) any row by a non-zero constant
(iii) adding (or subtracting) a constant multiple of one row to another row.

If any matrix A is transformed to another matrix B by a series of row operations, we say that A and B are equivalent matrices. More specifically we have.

Definition 3: A matrix B is said to be row-equivalent to a matrix A, if B can be obtained from A by a using a finite number of row operations.
Two linear systems $A\mathbf{x} = \mathbf{b}$ and $A'\mathbf{x} = \mathbf{b}'$ are said to be equivalent if they have the same solution. Hence, if a sequence of elementary operations on $A\mathbf{x} = \mathbf{b}$ produces the new system $A'\mathbf{x} = \mathbf{b}'$, then the systems $A\mathbf{x} = \mathbf{b}$ and $A'\mathbf{x} = \mathbf{b}'$ are equivalent.

Let us illustrate (Naive) Gauss elimination method by considering a system of three equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3. \end{aligned} \tag{1.3.1}$$

Let $a_{11} \neq 0$.. We multiply first equation of the system by $-\dfrac{a_{22}}{a_{11}}$ and add to the second equation . Then we multiply the first equation by $-\dfrac{a_{31}}{a_{11}}$ and add to the third equation. The new equivalent system (first derived system) then becomes

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 &= b_2^{(1)} \\ a_{32}^{(1)}x_3 + a_{33}^{(1)}x_3 &= b_3^{(1)} \end{aligned} \tag{1.3.2}$$

where

$$a_{22}^{(1)} = a_{22} - \frac{a_{21}}{a_{11}} \cdot a_{12}, \quad a_{23}^{(1)} = a_{23} - \frac{a_{21}}{a_{11}} \cdot a_{13},$$

$$b_2^{(1)} = b_2 - \frac{a_{21}}{a_{11}} \cdot b_1, \text{ etc.}$$

Next, we multiply the second equation of the derived system provided $a_{22}^{(1)} \neq 0$, by $-\dfrac{a_{32}^{(1)}}{a_{22}^{(1)}}$ and add to the third equation of (3.4.2). The system becomes

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 &= b_2^{(1)} \\ a_{33}^{(2)}x_3 &= b_3^{(2)} \end{aligned} \tag{1.3.3}$$

where

$$a_{33}^{(2)} = a_{33}^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}} \cdot a_{23}^{(1)}$$

and

$$b_3^{(2)} = b_3^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}} b_2^{(1)}.$$

This system is an upper-triangular system and can be solved using back substitutions method provided $a_{33}^{(2)} \neq 0$. That is, the last equation gives $x_3 = \dfrac{b_3^{(2)}}{a_{33}^{(2)}}$; then substituting this value of x_3 in the last but one equation (second) we get the value of x_2 and then substituting the obtained values of x_3 and x_2 in the first equation we compute x_1. This process of solving an upper-triangular system of linear equations in often called **back substitution**. We illustrate this by the following example:

Example 1: Solve the following system of equations consisting of four equations.

(Equation 1) E_1: $x_1 + x_2 + 0.x_3 + 3x_4 = 4$
E_2: $2x_1 + x_2 - x_3 + x_4 = 1$
E_3: $3x_1 - x_2 - x_3 + 2x_4 = -3$
E_4: $-x_1 + 2x_2 + 3x_3 - x_4 = 4.$

Solution: The first step is to use first equation to eliminate the unknown x_1 from second, third and fourth equation. This is accomplished by performing $E_2 - 2E_1$, $E_3 - 3E_1$ and $E_4 + E_1$. This gives the derived system as

$$\begin{aligned}
E'_1: &\quad x_1 + x_2 + 0.x_3 + 3x_4 = 4\\
E'_2: &\quad -x_2 - x_3 + 5x_4 = -7\\
E'_3: &\quad -4x_2 - x_3 - 7x_4 = -15\\
E'_4: &\quad 3x_2 + 3x_3 + 2x_4 = 8.
\end{aligned}$$

In this new system, E'_2 is used to eliminate x_2 from E'_3 and E'_4 by performing the operations $E'_3 - 4E'_2$ and $E'_4 + 3E'_2$. The resulting system is

$$\begin{aligned}
E''_1: &\quad x_1 + x_2 + 0.x_3 + 3x_4 = 4\\
E''_2: &\quad -x_2 - x_3 + 5x_4 = -7\\
E''_3: &\quad 3x_3 + 13x_4 = 13\\
E''_4: &\quad -13x_4 = -13.
\end{aligned}$$

This system of equation is now in triangular form and can be solved by back substitution. E''_4 gives $x_4 = 1$, E''_3 gives

$$x_3 = \frac{1}{3}(13 - 13x_4) = \frac{1}{3}(13 - 13 \times 1) = 0.$$

E''_2 gives $x_2 = -(-7 + 5x_4 + x_3) = -(-7 + 5 \times 1 + 0) = 2$ and E''_1 gives $x_1 = 4 - 3x_4 - x_2 = 4 - 3 \times 1 - 2 = -1$.

The above procedure can be carried out conveniently in matrix form as shown below:

We consider the Augmented matrix [A1**b**] and perform the elementary row operations on the augmented matrix.

$$[A \mid \mathbf{b}] \quad \left[\begin{array}{cccc|c} 1 & 2 & 0 & 3 & 4 \\ 2 & 1 & -1 & 1 & 1 \\ 3 & -1 & -1 & 2 & -3 \\ -1 & 2 & 3 & -1 & 4 \end{array}\right] \quad R_2 - 2R_1,\ R_3 - 3R_1 \text{ and } R_4 + R_1 \text{ gives}$$

$$= \quad \left[\begin{array}{cccc|c} 1 & 2 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & -4 & -1 & -7 & -15 \\ 0 & 3 & 3 & 2 & 8 \end{array}\right] \quad R_3 - 4R_2,\ R_4 + 3R_2 \text{ gives}$$

$$= \quad \left[\begin{array}{cccc|c} 1 & 2 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right]$$

This is the final equivalent system:

$$\begin{aligned}
x_1 + x_2 + 0x_3 + 3x_4 &= 4\\
-x_2 - x_3 - 5x_4 &= -7\\
3x_3 + 13x_4 &= 13\\
-13x_4 &= -13.
\end{aligned}$$

The method works with the assumption that none of the elements a_{11}, $a_{22}^{(1)}$, …, $a_{n-1,n-1}^{(n-2)}$, $a_{n,n}^{(n-1)}$ is zero. This does not necessarily mean that the linear system is not solvable, but the following technique may yield the solution:

Suppose $a_{kk}^{(k-1)} = 0$ for some $k = 2, \ldots, n-2$. The kth column of $(k-1)$th equivalent system from the kth row is searched for the first non zero entry. If $a_{pk}^{(k)} \neq 0$ for some p,

$k + 1 \le p \le n$, then interchange R_k by R_p to obtain an equivalent system and continue the procedure. If $a_{pk}^{(k)} = 0$ for p = k, k + 1, …, n, it can be shown that the linear system does not have a unique solution and hence the procedure is terminated.

You may now solve the following exercises:

E1) Solve the system of equations
$3x_1 + 2x_2 + x_3 = 3$
$2x_1 + x_2 + x_3 = 0$
$6x_1 + 2x_2 + 4x_3 = 6$
using Gauss elimination method. Does the solution exist?

E2) Solve the system of equations
$16x_1 + 22x_2 + 4x_3 = -2$
$4x_1 - 3x_2 + 2x_3 = 9$
$12x_1 + 25x_2 + 2x_3 = -11$
using Gauss elimination method and comment on the nature of the solution.

E3) Solve the system of equations by Gauss elimination.
$x_1 - x_2 + 2x_3 - x_4 = -8$
$2x_1 - 2x_2 + 3x_3 - 3x_4 = -20$
$x_1 + x_2 + x_3 + 0.x_4 = -2$
$x_1 - x_2 + 4x_3 + 3x_4 = 4$

E4) Solve the system of equations by Gauss elimination.
$x_1 + x_2 + x_3 + x_4 = 7$
$x_1 + x_2 + 0.x_3 + 2x_4 = 8$
$2x_1 + 2x_2 + 3x_3 + 0.x_4 = 10$
$-x_1 - x_2 - 2x_3 + 2x_4 = 0$

E5) Solve the system of equation by Gauss elimination.
$x_1 + x_2 + x_3 + x_4 = 7$
$x_1 + x_2 + 2x_4 = 5$
$2x_1 + 2x_2 + 3x_3 = 10$
$-x_1 - x_2 - 2x_3 + 2x_4 = 0$

It can be shown that in Gauss elimination procedure and back substitution $(2n^3 + 3n^2 - 5n)/6 + \frac{n^2 + n}{2}$ multiplications/divisions and $\frac{n^3 - n}{3} + \frac{n^2 - n}{2}$ additions/subtractions are performed respectively. The total arithmetic operation involved in this method of solving a n × n linear system is $\frac{n^3 + 3n^2 - n}{3}$ multiplication/divisions and $\frac{2n^3 + 3n^n - 5n}{6}$ additions/subtractions.

Definition 4: In Gauss elimination procedure, the diagonal elements $a_{11}, a_{22}^{(1)}, a_{33}^{(2)}$, which have been used as divisors are called pivots and the corresponding equations, are called pivotal equations.

1.3.3 Pivoting Strategies

If at any stage of the Gauss elimination, one of these pivots say $a_{ii}^{i-1} (a_{11}^{(0)} = a_{11})$, vanishes then we have indicated a modified procedure. But it may also happen that the pivot $a_{ii}^{(i-1)}$, though not zero, may be very small in magnitude compared to the

remaining elements (≥ i) in the ith column. Using a small number as divisor may lead to growth of the round-off error. The use of large multipliers like

$$\frac{-a_{i+1}^{(i-12)}}{a_{ii}^{(i-1)}}, i \quad \frac{a_{i+2,i}^{(i-1)}}{a_{ii}^{(i-1)}}$$

etc. will lend to magnification of errors both during the elimination phase and during the back substitution phase of the solution procedure. This can be avoided by rearranging the remaining rows (from ith row up to nth row) so as to obtain a non-vanishing pivot or to choose one that is largest in magnitude in that column. This is called pivoting strategy.

There are two types of pivoting strategies: partial pivoting (maximal column pivoting) and complete pivoting. We shall confine to simple partial pivoting and complete pivoting. That is, the method of scaled partial pivoting will not be discussed. Also there is a convenient way of carrying out the pivoting procedure where instead of interchanging the equations all the time, the n original equations and the various changes made in them can be recorded in a systematic way using the augmented matrix [A1**b**] and storing the multiplies and maintaining pivotal vector. We shall just illustrate this with the help of an example. However, leaving aside the complexities of notations, the procedure is useful in computation of the solution of a linear system of equations.

If exact arithmetic is used throughout the computation, pivoting is not necessary unless the pivot vanishes. But, if computation is carried up to a fixed number of digits (precision fixed), we get accurate results if pivoting is used.

The following example illustrates the effect of round-off error while performing Gauss elimination:

Example 2: Solve by the Gauss elimination the following system using four-digit arithmetic with rounding.

$$\begin{aligned} 0.003000x_1 + 59.14\ x_2 &= 59.17 \\ 5.291x_1 - \ 6.130x_2 &= 46.78. \end{aligned}$$

Solution: The first pivot element $a_{11}^0 = a_{11} = 0.0030$ and its associated multiplier is

$$\frac{5.291}{0.0030} = 1763.66 \approx 1763$$

Performing the operation of elimination of x_1 from the second equation with appropriate rounding we got

$$\begin{aligned} 0.003000\ x_1 + \ 59.14x_2 &= \ \ 59.17 \\ -\ 104300\ x_2 &= -104400 \end{aligned}$$

By backward substitution we have

$$x_2 = 1.001 \text{ and } x_1 = \frac{59.17 - (59.14)(1.001)}{0.00300} = -10.0$$

The linear system has the exact solution $x_1 = 10.00$ and $x_2 = 1{,}000$.

However, if we use the second equation as the first pivotal equation and solve the system, the four digit arithmetic with rounding yields solution as $x_1 = 10.00$ and $x_2 = 1.000$. This brings out the importance of partial or maximal column pivoting.

Partial pivoting (Column Pivoting)

In the first stage of elimination, instead of using $a_{11} \neq 0$ as the pivot element, the first column of the matrix A ([A1**b**]) is searched for the largest element in magnitude and this largest element is then brought at the position of the first pivot by interchanging first row with the row having the largest element in magnitude in the first column.

Next, after elimination of x_1, the second column of the derived system is searched for the largest element in magnitude among the (n – 1) element leaving the first element. Then this largest element in magnitude is brought at the position of the second pivot by interchanging the second row with the row having the largest element in the second column of the derived system. The process of searching and interchanging is repeated in all the (n – 1) stages of elimination. For selecting the pivot we have the following algorithm:

For i = 1, 2, …, n find j such that

$$\left|a_{ji}^{(i-1)}\right| = \max_{i \le k \le n} \left|a_{ki}^{(i-1)}\right| \qquad \left(a_{ji}^{0} = a_{ji}\right)$$

Interchange ith and jth rows and eliminate x_i.

Complete Pivoting

In the first stage of elimination, we look for the largest element in magnitude in the entire matrix A first. If the element is a_{pq}, then we interchange first row with pth row and interchange first column with qth column, so that a_{pq} can be used as a first pivot. After eliminating x_q, the process is repeated in the derived system, more specifically in the square matrix of order n – 1, leaving the first row and first column. Obviously, complete pivoting is quite cumbersome.

Scaled partial pivoting (Sealed column pivoting)

First a scale factor d_i for each row i is defined by $d_i = \max_{i \le j \le n} \left|a_{ij}\right|$

If $d_i = 0$ for any i, there is no unique solution and procedure is terminated. In the first stage choose the first integer k such that

$$\left|a_{k1}\right| / d_k = \max_{i \le j \le n} \left|a_{j1} / d_j\right|$$

interchange first row and kth row and eliminate x_1. The process is repeated in the derived system leaving aside first row and first column.

We now illustrate these pivoting strategies in the following examples.

Example 3: Solve the following system of linear equations with partial pivoting

$$\begin{aligned} x_1 - x_2 + 3x_3 &= 3 \\ 2x_1 + x_2 + 4x_3 &= 7 \\ 3x_1 + 5x_2 - 2x_3 &= 6 \end{aligned}$$

$$[A1b] = \left(\begin{array}{ccc|c} 1 & -1 & 3 & 3 \\ 2 & 1 & 4 & 7 \\ 3 & 5 & -2 & 6 \end{array}\right) \quad R_1 - \frac{1}{3}R_3, R_2 - \frac{2}{3}R_3$$

$$= \left(\begin{array}{ccc|c} 0 & -\frac{8}{3} & \frac{11}{3} & 1 \\ 0 & -\frac{7}{3} & \frac{16}{3} & 3 \\ 3 & 5 & -2 & 6 \end{array}\right) \qquad R_2 - \frac{7}{3}\cdot\frac{3}{8}R_1$$

$$\left(\begin{array}{ccc|c} 0 & -\frac{8}{3} & \frac{11}{3} & 1 \\ 0 & 0 & \frac{51}{24} & \frac{17}{8} \\ 3 & 5 & -2 & 6 \end{array}\right)$$

Re-arranging the equations (3rd equation becomes the first equation and first equation becomes the second equation in the derived system), we have

$$3x_1 + 5x_2 - 2x_3 = 6$$
$$-\frac{8}{3}x_2 + \frac{11}{3}x_3 = 1$$
$$\frac{51}{24}x_3 = \frac{17}{18}$$

Using back substitution we have $x_1 = 1$, $x_2 = 1$ and $x_3 = 1$.

You may now solve the following exercises:

E6) Solve the system of linear equation given in the Example 3 by complete pivoting.

E7) Solve the system of linear equation given in Example 3 by scaled partial pivoting.

E8) Solve the system of equations with partial (maximal column) pivoting.

$$x_1 + x_2 + x_3 = 6$$
$$3x_1 + 3x_2 + 4x_3 = 20$$
$$2x_1 + x_2 + 3x_3 = 13$$

1.4 ITERATIVE METHODS

Consider the system of equations

$$A\mathbf{x} = \mathbf{b} \qquad \ldots (1.4.1)$$

Where A is an n × n non-singular matrix. An iterative technique to solve the n x n linear system (1.4.1) starts with an initial approximation $\underline{x}^{(0)}$ to the solution **x**, and generates a sequence of vectors $\{\underline{x}^k$ that **converges** to **x**, the actual solution vector (When $\max_{1\le i\le n}\left|x_i^{(k)} - x_i\right| < \varepsilon$ for some k when ε is a given small positive numbers.).

Most of these iterative techniques entails a process that converts the system A**x** = **b** into an equivalent system of the form **x** = T**x** + **c** for some n × n matrix T and vector **c**. In general we can write the iteration method for solving the linear system (3.5.1) in the form

$$x^{(k+1)} = Tx^{(k)} + c \qquad k = 0, 1, 2, \ldots,$$

T is called the iteration matrix and depends on A, **c** is a column vector which depends on A and **b**. We illustrate this by the following example.

Iterative methods are generally used when the system is large (when n > 50) and the matrix is sparse (matrices with very few non-zero entries).

Example 4: Convert the following linear system of equations into equivalent form $\mathbf{x} = T\mathbf{x} + \mathbf{c}$.

$$\begin{aligned} 10x_1 - x_2 + 2x_3 &= 6 \\ -x_1 + 11x_2 - x_3 + 3x_4 &= 25 \\ 2x_1 - x_2 + 10x_3 - x_4 &= -11 \\ 3x_2 - x_3 + 8x_4 &= 15 \end{aligned}$$

Solution: We solve the ith equation for x_i (assuming that $a_{ii} \neq 0 \ \forall$ i. If not, we can interchange equations so that is possible)

$$x_1 = \frac{1}{10}x_2 - \frac{1}{5}x_3 + \frac{3}{5}$$

$$x_2 = \frac{1}{11}x_1 + \frac{1}{11}x_3 - \frac{3}{11}x_4 + \frac{25}{11}$$

$$x_3 = -\frac{1}{5}x_1 + \frac{1}{10}x_2 + \frac{1}{10}x_4 - \frac{11}{10}$$

$$x_4 = -\frac{3}{8}x_2 + \frac{1}{8}x_3 + \frac{15}{8}$$

Here $T = \begin{bmatrix} 0 & \frac{1}{10} & -\frac{1}{5} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11} \\ -\frac{1}{5} & \frac{1}{10} & 0 & \frac{1}{10} \\ 0 & -\frac{3}{8} & \frac{1}{8} & 0 \end{bmatrix}$ and $\mathbf{c} = \begin{bmatrix} \frac{3}{5} \\ \frac{25}{11} \\ -\frac{11}{10} \\ \frac{15}{8} \end{bmatrix}$

1.4.1 The Jacobi Iterative Method

This method consists of solving the ith equation of $A\mathbf{x} = \mathbf{b}$ for x_i, to obtain

$$x_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(\frac{-a_{ij}x_j}{a_{ii}} \right) + \frac{b_i}{a_{ii}} \quad \text{for } i = 1, 2, \ldots, n$$

provided $a_{ii} \neq 0$.

We generate $\mathbf{x}^{(k+1)}$ from $\mathbf{x}^{(k)}$ for $k \geq 0$ by

$$x_i^{(k+1)} = \sum_{\substack{j=i \\ j \neq i}}^{n} \left(-a_{ij}x_j^{(k)} + b_i \right) \Big/ a_{ii} \quad i = 1,2,\ldots,n \qquad (1.4.2)$$

We state below a sufficient condition for convergence of the Jacobi Method.

Theorem

If the matrix A is strictly diagonally dominant, that is, if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, 2, \ldots n.$$

then the Jacobi iteration method (3.5.2) converges for any initial approximation $\mathbf{x}^{(0)}$.

Generally $\mathbf{x}^{(0)} = \mathbf{0}$ is taken in the absence of any better initial approximation.

Example 5: Solve the linear system $A\mathbf{x} = \mathbf{b}$ given in previous example (Example 4) by Jacobi method rounded to four decimal places.

Solution: Letting $\mathbf{x}^{(0)} = (0,0,0,0)^T$, we get

$$\mathbf{x}^{(1)} = (0.6000, 2.2727, -1.1000, 1.8750)^T$$

$$\mathbf{x}^{(2)} = (1.0473, 1.7159, -0.8052, 0.8852)^T \text{ and}$$

$$\mathbf{x}^{(3)} = (0.9326, 2.0533, -1.0493, 1.1309)^T$$

Proceeding similarly one can obtain

$$\mathbf{x}^{(5)} = (0.9890, 2.0114, -1.0103, 1.0214)^T \text{ and}$$

$$\mathbf{x}^{(10)} = (1.0001, 1.9998, -0.9998, 0.9998)^T.$$

The solution is $\mathbf{x} = (1, 2, -1, 1)^T$. You may note that $\mathbf{x}^{(10)}$ is a good approximation to the exact solution compared to $\mathbf{x}^{(5)}$.

You also observe that A is strictly diagonally dominant
(since $10 > 1 + 2$, $11 > 1 + 1 + 3$ $10 > 2 + 1 + 1$ and $8 > 3 + 1$).

Now we see how $A\mathbf{x} = \mathbf{b}$ is transformed to an equivalent system $\mathbf{x} = T\mathbf{x} + \mathbf{c}$.

The matrix can be written as

$A = D + L + U$ where

$$D = \begin{bmatrix} a_{11}, & 0 & \ldots & 0 \\ 0 & a_{22}, & \ldots & 0 \\ 0 & \ldots & 0, & a_{nn} \end{bmatrix}$$

$$L = \begin{bmatrix} 0 & 0 & \ldots & \ldots & 0 \\ a_2 & 0 & \ldots & \ldots & 0 \\ a_3, & a_{32} & 0 & \ldots & 0 \\ a_n, & a_{n2} & \ldots & a_{n,\,n-1} & \end{bmatrix}$$

$$U = \begin{bmatrix} 0 & a_{12} & & \ldots & a_{1n} \\ 0 & 0 & a_{23} & \ldots & a_{2n} \\ 0 & 0 & 0 & \ldots & a_{n-1,\,n} \end{bmatrix}$$

Since $(D + L + U)\,\mathbf{x} = \mathbf{b}$

$$D\mathbf{x} = -(L + U)\,\mathbf{x} + \mathbf{b}$$

$$\mathbf{x} = -D^{-1}(L + U)\,\mathbf{x} + D^{-1}\mathbf{b}$$

i.e. $T = D^{-1}(L + U)$ and $\mathbf{c} = D^{-1}\mathbf{b}$.

In Jacobi method, each of the equations is simultaneously changed by using the most recent set of **x**-values. Hence the Jacobi method is called method of simultaneous displacements.

You may now solve the following exercises:

E9) Perform five iterations of the Jacobi method for solving the system of equations.

$$\begin{bmatrix} 5 & -1 & -1 & -1 \\ -1 & 10 & -1 & -1 \\ -1 & -1 & 5 & -1 \\ -1 & -1 & -1 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -4 \\ 12 \\ 8 \\ 34 \end{bmatrix}$$

Starting with $\mathbf{x}^{(0)} = (0,0,0,0)$. The exact solution is $\mathbf{x} = (1,2,3,4)T^n$. How good $\mathbf{x}^{(5)}$ as an approximation to $\mathbf{x}$?

E10) Perform four iterations of the Jacobi method for solving the following system of equations.

$$\begin{bmatrix} 2 & -1 & -0 & -0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

With $\mathbf{x}^{(0)} = (0.5, 0.5, 0.5, 0.5)^T$. Here $\mathbf{x} = (1, 1, 1, 1)^T$. How good $\mathbf{x}^{(5)}$ as an approximation to $\mathbf{x}$?

1.4.2 The Gauss-Seidel Iteration Method

In this method, we can write the iterative scheme of the system of equations $A\mathbf{x} = \mathbf{b}$ as follows:

$$a_{11}x_1^{(k+1)} = -a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \ \ldots \ - a_n x_n^{(k)} + b_1$$

$$a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} = -a_{23}x_3^{(k)} - \ \ldots \ - a_{2n}x_n^{(k)} + b_2$$

.

.

$$a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} \ldots + a_{nn}x_n^{(k+1)} = + b_n$$

In matrix form, this system can be written as $(D + L)\ \mathbf{x}^{(k+1)} = -U\mathbf{x}^{(k)} + \mathbf{b}$ with the same notation as adopted in Jacobi method.

From the above, we get

$$x^{(k+1)} = -(D+L)^{-1}Ux^{(k)} + (D+L)^{-1}b$$

$$= T\boldsymbol{x}^{(k)} + \boldsymbol{c}_n$$

i.e. $T = -(D+L)^{-1n}\,U$ and $\mathbf{c} = (D + L)^{-1}\mathbf{b}$

This iteration method is also known as the method of successive displacement.

For computation point of view, we rewrite (A x) as

$$x_i^{(k+1)} = -\frac{1}{a_{ii}}\left[\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=i+1}^{x} a_{ij}x_j^{(k)} - b_i\right]$$

$i = 1, 2, \ldots, n$

Also in this case, if A is diagonally dominant, then iteration method always converges. In general Gauss-Seidel method will converge if the Jacobi method converges and will converge at a faster rate. You can observe this in the following example. We have not considered the problem: How many iterations are needed to have a reasonably good approximation to **x**? This needs the concept of matrix norm.

Example 6: Solve the linear system A**x** = **b** given in Example 4 by Gauss-Seidel method rounded to four decimal places. The equations can be written as follows:

$$
\begin{aligned}
x_1^{(k+1)} &= \frac{1}{10}x_2^{(k)} - \frac{1}{3}x_3^{(k)} + \frac{3}{5} \\
x_2^{(k+1)} &= \frac{1}{11}x_1^{(k+1)} + \frac{1}{11}x_3^{k} - \frac{3}{11}x_4^{(k)} + \frac{25}{11} \\
x_3^{(k+1)} &= -\frac{1}{3}x_1^{(k+1)} + \frac{1}{10}x_2^{(k+1)} + \frac{1}{10}x_4^{(k)} - \frac{11}{10} \\
x_4^{(k+1)} &= -\frac{3}{8}x_2^{(k+1)} + \frac{1}{8}x_3^{(k+1)} + \frac{15}{8}.
\end{aligned}
$$

Letting $\mathbf{x}^{(0)} = (0, 0, 0, 0)^T$ we have from first equation

$$
\begin{aligned}
x_1^{(1)} &= 0.6000 \\
x_2^{(1)} &= \frac{0.6000}{3} + \frac{25}{11} = 2.3273 \\
x_3^{(1)} &= -\frac{0.6000}{3} + \frac{1}{10}(2.3273) - \frac{11}{10} = -0.1200 + 0.2327 - 1.1000 = -0.9873 \\
x_4^{(1)} &= -\frac{3}{8}(2.3273) + \frac{1}{8}(-0.9873) + \frac{15}{8}
\end{aligned}
$$

$= -0.8727 - 0.1234 + 1.8750$

$= 0.8789$

Using $\mathbf{x}^{(1)}$ we get

$$\mathbf{x}^{(2)} = (1.0300, 2.037, -1.014, 0.9844)^T$$

and we can check that

$$\mathbf{x}^{(5)} = (1.0001, 2.0000, -1.0000, 1.0000)^T$$

Note that $\mathbf{x}^{(5)}$ is a good approximation to the exact solution. Here are a few exercises for you to solve.

You may now solve the following exercises:

E11) Perform four iterations (rounded to four decimal places) using Jacobi Method and Gauss-Seidel method for the following system of equations.

$$
\begin{bmatrix} -8 & 1 & 1 \\ 1 & -5 & -1 \\ 1 & 1 & -4 \end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} =
\begin{bmatrix} 1 \\ 16 \\ 7 \end{bmatrix}
$$

With $\mathbf{x}^{(0)} = (0, 0, 0)^T$. The exact solution is $(-1, -4, -3)^T$. Which method gives better approximation to the exact solution?

E12) For linear system given in E10), use the Gauss Seidel method for solving the system starting with $\mathbf{x}^{(0)} = (0.5, 0.5, 0.5, 0.5)^T$ obtain $\mathbf{x}^{(4)}$ by Gauss-Seidel method and compare this with $\mathbf{x}^{(4)}$ obtained by Jacobi method in E10).

1.4.3 Comparison of Direct and Iterative Methods

Both the methods have their strengths and weaknesses and a choice is based on the particular linear system to be solved. We mention a few of these below:

Direct Method

1. The direct methods are generally used when the matrix A is dense or filled, that is, there are few zero elements, and the order of the matrix is not very large, say n < 50.

2. The rounding errors may become quite large for ill conditioned equations (If at any stage during the application of pivoting strategy, it is found that all values of $\{ |a_{mk}\ 1|$ for m = k + 1, to n are less than a pre-assigned small quantity ε, then the equations are ill-conditioned and no useful solution is obtained). Ill-conditioned matrices are not discussed in this unit.

Iterative Method

1. These methods are generally used when the matrix A is sparse and the order of the matrix A is very large say n > 50. Sparse matrices have very few non-zero elements.

2. An important advantage of the iterative methods is the small rounding error. Thus, these methods are good choice for ill-conditioned systems.

3. However, convergence may be guaranteed only under special conditions. But when convergence is assured, this method is better than direct.

With this we conclude this unit. Let us now recollect the main points discussed in this unit.

1.5 SUMMARY

In this unit we have dealt with the following:

1. We have discussed the direct methods and the iterative techniques for solving linear system of equations A**x** = **b** where A is an n × n, non-singular matrix.

2. The direct methods produce the exact solution in a finite number of steps provided there are no round off errors. Direct method is used for linear system A**x** = **b** where the matrix A is dense and order of the matrix is less than 50.

3. In direct methods, we have discussed Gauss elimination, and Gauss elimination with partial (maximal column) pivoting and complete or total pivoting.

4. We have discussed two iterative methods, Jacobi method and Gauss-Seidel method and stated the convergence criterion for the iteration scheme. The iterative methods are suitable for solving linear systems when the matrix is sparse and the order of the matrix is greater than 50.

1.6 SOLUTION/ANSWERS

E1) $[A1\mathbf{b}]$ $\left[\begin{array}{ccc|c} -3 & 2 & 1 & 3 \\ 2 & 1 & -1 & 0 \\ 6 & 2 & 4 & 6 \end{array}\right]$

$a_{11} \neq 0 \longrightarrow$ $\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & \frac{1}{3} & -\frac{1}{3} & -2 \\ 0 & -2 & 2 & 0 \end{array}\right]$

$a_{22}^{(1)} \neq 0 \longrightarrow$ $\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & -\frac{1}{3} & -2 \\ 0 & 0 & 0 & 12 \end{array}\right]$

This system has no solution since x_3 cannot be determined from the last equation. This system is said to be inconsistent. Also note that del (A) = 0.

E2) $[A1\mathbf{b}]$ $\left[\begin{array}{ccc|c} 16 & 22 & 4 & -2 \\ 4 & -3 & 2 & 9 \\ 12 & 25 & 2 & -11 \end{array}\right]$

$a_{11} \neq 0 \longrightarrow$ $\left[\begin{array}{ccc|c} 16 & 22 & 4 & 2 \\ 0 & -\frac{17}{2} & 1 & \frac{19}{2} \\ 0 & \frac{17}{2} & -1 & -\frac{19}{2} \end{array}\right]$

$a_{22}^{(1)} \neq 0 \longrightarrow$ $\left[\begin{array}{ccc|c} 16 & 22 & 4 & 2 \\ 0 & -\frac{17}{2} & 1 & \frac{19}{2} \\ 0 & 0 & 0 & 0 \end{array}\right]$ $\Rightarrow x_3$ = arbitrary value, $x_2 = \frac{-2}{17}\left(\frac{19}{2} - x_3\right)$ and $x_3 = \frac{1}{6}(-2 - 22x_3 - 22x_3)$

This system has infinitely many solutions. Also you may check that det (A) = 0.

E3) Final derived system:

$\left[\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 0 & 2 & 4 \end{array}\right]$ and the solution is $x_4 = 2$, $x_3 = 2$ $x_2 = 3$, $x_1 = -7$.

E4) Final derived system:

$$\left[\begin{array}{cccc|c} 1 & -1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right]$$ and the solution is

$x_4 = 3$, $x_3 = 2$, x_2 arbitrary and $x_1 = 2 - x_2$.

Thus this linear system has infinite number of solutions.

E5) Final derived system:

$$\left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & -2 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right]$$ and the solution does not

exist since we have $x_4 = 3$, $x_3 = 2$ and third equation ϕ $-x_3 + x_4 = -2$ implies $1 = -2$, leading to a contradiction.

E6) Since $|a_{32}|$ is maximum we rewrite the system as

$$\left[\begin{array}{ccc|c} 5 & 3 & -2 & 6 \\ 1 & 2 & 4 & 7 \\ -1 & 1 & 3 & 3 \end{array}\right]$$ by interchanging R_1 and R_3 and C_1 and C_2

$R_2 - \frac{1}{5}R_1, R_3 + \frac{1}{5}R_1$ gives

$$\left[\begin{array}{ccc|c} 5 & 3 & -2 & 6 \\ 0 & \frac{7}{5} & \frac{22}{5} & \frac{29}{5} \\ 0 & \frac{8}{5} & \frac{13}{5} & \frac{21}{5} \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 5 & -2 & 3 & 6 \\ 0 & \frac{22}{5} & \frac{7}{5} & \frac{29}{5} \\ 0 & \frac{13}{5} & \frac{8}{5} & \frac{21}{5} \end{array}\right]$$ by interchanging C_2 and C_3

Since $|a_{23}|$ is maximum –

By $R_3 - \frac{5}{12} x \frac{13}{15} R_2$ we have

$$\left[\begin{array}{ccc|c} 5 & -2 & 3 & 6 \\ 0 & \frac{22}{5} & \frac{7}{5} & \frac{29}{5} \\ 0 & 0 & \frac{17}{22} & \frac{43}{22} \end{array}\right]$$

$$5x_2 + 3x_1 - 2x_3 = 6$$

$$\frac{22}{5}x_3 + \frac{7}{5}x_2 = \frac{29}{5}$$

$$\frac{17}{22}x_2 = \frac{17}{22}$$

We have $x_2 = 1$, $\frac{22}{5}x_3 = \frac{29}{5} - \frac{7}{5} = \frac{22}{5} \Rightarrow x_3 = 1$

$3x_1 = 6 - 5 + 2 \Rightarrow x_1 = 1$

E7) For solving the linear system by sealed partial pivoting we note that $d_1 = 3$, $d_2 = 4$ and $d_3 = 5$ in

$$W = [A1\mathbf{b}] = \begin{bmatrix} 1 & -1 & 3 & 3 \\ 2 & 1 & 4 & 7 \\ 3 & 5 & -2 & 6 \end{bmatrix} \quad p = [a, 2, 3]^T$$

Since $\max\left\{\frac{1}{3}, \frac{2}{4}, \frac{3}{5}\right\} = \frac{3}{5}$, third equation is chosen as the first pivotal equation.

Eliminating x_1 we have

$$\begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix} \quad W^1 = \begin{bmatrix} \frac{1}{3} & \frac{8}{3} & \frac{11}{3} & 1 \\ -\frac{2}{3} & -\frac{7}{3} & \frac{16}{3} & 3 \\ 3 & 5 & -2 & 6 \end{bmatrix}$$

where we have used a square to enclose the pivot element and

in place of zero entries, after elimination of x_1 from 1st and 2nd equation, we have stored the multipliers. Here $m_{11} = \frac{a_{11}}{a_{31}} = \frac{1}{3}$ *and* $m_{2,1} = \frac{a_{21}}{a_{31}} = \frac{2}{3}$

Instead of interchanging rows (here R_1 and R_3)we keep track of the pivotal equations being used by the vector $p=[3,2,1]^T$

In the next step we consider $\max\left\{\frac{7}{3} \cdot \frac{1}{4}\frac{8}{3} \cdot \frac{1}{3}\right\} = \frac{8}{9}$

So the second pivotal equation is the first equation.

i.e. $p = [3, 1, 3]^T$ and multiplier is $-\frac{7}{3} - \frac{8}{3} = \frac{7}{8} = m_{2,2}$

$$\text{and } W^{(2)} = \begin{bmatrix} -\frac{8}{3} & \frac{11}{3} & 1 & \\ \frac{7}{8} & \frac{17}{8} & \frac{17}{8} & p = [3, 1, 2]^T \\ 5 & -2 & 6 & \end{bmatrix}$$

The triangular system is as follows:

$$3x_1 + 5x_2 - 2x_3 = 6$$
$$-\frac{8}{3}x_2 + \frac{11}{3}x_3 = 1$$
$$\frac{17}{8}x_3 = \frac{17}{8}$$

By back substitution, this yields $x_1 = 1$, $x_2 = 1$ and $x_3 = 1$.

Remark: The **p** vector and storing of multipliers help solving the system A**x** = **b**' where b is changed **b**'.

E8) $[A, \mathbf{b}] = \begin{bmatrix} 1 & 1 & 1 & 6 \\ 3 & 3 & 4 & 20 \\ 2 & 1 & 3 & 13 \end{bmatrix} \rightarrow \begin{bmatrix} 3 & 3 & 4 & 20 \\ 1 & 1 & 1 & 6 \\ 2 & 1 & 3 & 13 \end{bmatrix}$

$$\xrightarrow{R_2 - \frac{1}{3}R_1, R_3 - \frac{2}{3}R_1} \begin{bmatrix} 3 & 3 & 4 & 20 \\ 0 & 0 & -\frac{1}{3} & -\frac{2}{3} \\ 0 & -1 & \frac{1}{3} & -\frac{1}{3} \end{bmatrix} \quad \begin{bmatrix} 13 & 3 & 4 & 20 \\ 0 & -1 & \frac{1}{3} & -\frac{1}{3} \\ 0 & 0 & -\frac{1}{3} & -\frac{2}{3} \end{bmatrix}$$

Since the resultant matrix is in triangular form, using back substitution we get $x_3 = 2$, $x_2 = 1$ and $x_1 = 3$.

E9) Using $\mathbf{x}^{(0)} = [0, 0, 0, 0]^T$ we have

$\mathbf{x}^{(1)} = [-0.8, 1.2, 1.6, 3.4]^T$

$\mathbf{x}^{(2)} = [0.44, 1.62, 2.36, 3.6]^T$

$\mathbf{x}^{(3)} = [0.716, 1.84, 2.732, 3.842]^T$

$\mathbf{x}^{(4)} = [0.8823, 1.9290, 2.8796, 3.9288]^T$

E10) Using $\mathbf{x}^{(0)} = [0.5, 0.5, 0.5, 0.5]^T$, we have

$\mathbf{x}^{(1)} = [0.75, 0.5, 0.5, 0.75]^T$

$\mathbf{x}^{(2)} = [0.75, 0.625, 0.625, 0.75]^T$

$\mathbf{x}^{(3)} = [0.8125, 0.6875, 0.6875, 0.8125]^T$

$\mathbf{x}^{(4)} = [0.8438, 0.75, 0.75, 0.8438]^T$

E 11) By Jacobi method we have

$\mathbf{x}^{(1)} = [-0.125, -3.2, -1.75]^T$

$\mathbf{x}^{(2)} = [-0.7438, -3.5750, -2.5813]^T$

$\mathbf{x}^{(3)} = [-0.8945, -3.8650, -2.8297]^T$

$\mathbf{x}^{(4)} = [-0.9618, -3.9448, -2.9399]$

Where as by Gauss-Seidel method, we have

$\mathbf{x}^{(1)} = [-0.125, -3.225, -2.5875]^T$

$\mathbf{x}^{(2)} = [-0.8516, -3.8878, -2.9349]^T$

$\mathbf{x}^{(3)} = [-0.9778, -3.9825, -2.9901]^T$

$\mathbf{x}^{(4)} = [-0.9966, -3.9973, -2.9985]^T$

E12) Starting with the initial approximation

$\mathbf{x}^{(0)} = [0.5, 0.5, 0.5, 0.5]^T$, we have the following iterates:

$\mathbf{x}^{(1)} = [0.75, 0.625, 0.5625, 0.7813]^T$

$\mathbf{x}^{(2)} = [0.8125, 0.6875, 0.7344, 0.8672]^T$

$$\mathbf{x}^{(3)} = [0.8438, 0.7891, 0.8282, 0.9141]^T$$

$$\mathbf{x}^{(4)} = [0.8946, 0.8614, 0.8878, 0.9439]^T$$

Since the exact solution is $\mathbf{x} = [1, 1, 1, 1]^T$, the Gauss, Seidel method gives better approximation than the Jacobi method at fourth iteration.

UNIT 2 INTERPOLATION

Structure **Page Nos.**

2.0 INTRODUCTION

The interpolation has been defined as the art of reading between the lines of a table, and in elementary mathematics the term usually denotes the process of computing intermediate values of a function from a set of given values of that function. Suppose the value of the function f(x) (instead of the analytical formula representing the function) is tabulated at a discrete set of values of the argument x at $x_0, x_1, \ldots, x_n$ $x_0 \leq x_1 \leq \ldots \leq x_n$. If the value of f(x) is to be found at some point ξ in the interval $[x_o, x_n]$ and ξ is not one of the nodes x_i, then value is estimated by using the known values of f(x) at the surrounding points. Such estimates of f(x) can be made using a function that fits the given data. If the point ξ is outside the interval $[x_o, x_n]$, then the estimation of $f(\xi)$ is called extrapolation.

The general problem of interpolation, however, is much more complex than this. In higher mathematics we often have to deal with functions whose analytical form is either totally unknown or else is of such a nature, complicated or otherwise, that the function cannot easily be subjected to certain operations like differentiation and integration etc. In either case, it is desirable to replace the given function by another which can be more readily handled.

We derive various forms of the interpolating polynomial. Polynomials are used as the basic means of approximation in nearly all areas of numerical analysis. We have divided our discussion on polynomial interpolation into 3 sections. First we discuss Lagrange form of interpolating polynomial for unequally spaced nodes. Also general expression for the error (truncation) of polynomial interpolation is obtained which provides the estimates of the error in polynomial approximation. In next section, we deal with another very useful form of interpolating polynomial called the Newton form of interpolating polynomial. Also we obtain another expression for the error term in term of divided difference.

Finally we deal with some useful forms of interpolating polynomial for equally spaced nodes like Newton's forward difference form Newton's backward difference form after introducing the concepts of forward and backward differences.

2.1 OBJECTIVES

After going through this unit, you will be able to:

- find the Lagrange's form and Newton's divided difference form of interpolating polynomials interpolating f(x) at n + 1 distinct nodal points;
- compute the approximate value of f a non-tabular point;
- compute the value of $\bar{x}$ (approximately) given a number $\bar{y}$ such that f($\bar{x}$) = $\bar{y}$ (inverse interpolation);
- compute the error committed in interpolation, if the function is known, at a non-tabular point of interest;
- find an upper bound in the magnitude of the error;
- form a table of divided differences and find divided differences with a given set of arguments from the table;
- write a forward (backward) difference in terms of function values from a table of forward (backward) differences and locate a difference of given order at a given point from the table; and
- obtain the interpolating polynomial of f(x) for a given data by Newton's forward (backward) difference formula.

2.2 LAGRANGE'S FORM

2.2.1 Problem of Interpolation

We are here concerned with a real-valued function f(x) defined on the interval [a, b] such that the analytical formula representing the function is unknown, but the values of the function f(x) are given for a given set of n + 1 distinct values of $x = x_0, x_1, x_2, \ldots, x_n$ where $x_0 < x_1 < x_2, \ldots, < x_n$ lying in the interval [a, b]. We denote $f(x_k)$ by f_k. The technique of determining an approximate value of f(x) for a non-tabular value of x which lies in the interval is called **interpolation**. The process of determining the value of f(x) for a value of x lying outside the interval [a, b] is called **extrapolation**. However, we shall assume that f(x) is defined in $(-\infty, \infty)$ in which it is continuously differentiable a sufficient number of times.

2.2.2 Lagrange's Form of Interpolating Polynomial

In this section, we derive a polynomial P(x) of degree ≤ n which agrees with values of f at the given (n + 1) distinct points, called the nodes or abscissas. In other words, we can find a polynomial P(x) such that $P(x_i) = f_i$, i = 0, 1, 2, …, n. Such a polynomial P(x) is called the interpolating polynomial of f(x).

First we prove the existence of an interpolating polynomial by actually constructing one such polynomial having the desired property. The uniqueness of the interpolating polynomial is proved by invoking the corollary of the fundamental theorem of algebra. Then we derive a general expression for error in approximating the function by the interpolating polynomial at a point and this allows us to calculate a bound on the error over an interval.

In polynomial interpolation the approximating function is taken to be a polynomial $P_n(x)$ of degree ≤ n such that

$$P_n(x_i) = f(x_i) \qquad i = 0,1,2 \ldots..n \qquad (2.2.1)$$

Let $P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x x + a_0$. Then, by conditions (4.3.1) we have

$a_n x_i^n + a_{n-1} x_i^{n-1} + ... + a_1 x_i + a_0 = f(x_i), \quad i = 0, 1, 2\ldots, n.$

This is a system of n + 1 linear equations in n + 1 unknowns $a_0, a_1, \ldots, a_n$. Since the determinant of the co-efficients

$$\begin{vmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \ldots & & \ldots & \\ 1 & x_n & \ldots & x_n^n \end{vmatrix} = \prod_{i>j} (x_i - x_j) \neq 0$$

as $x_o, x_1 \ldots\ldots x_n$ are distinct points/nodes, the values of $a_0, a_1, \ldots, a_n$ can be uniquely determined so that the polynomial $P_n(x)$ exists and is unique. But this does not give us the explicit form of $P_n(x)$. Hence in the following we first show the existence of $P_n(x)$ by constructing one such polynomial and then prove its uniqueness.

Theorem 1: Let $x_0, x_1, \ldots, x_n$ be n + 1 distinct points on the real line and let f(x) be a real-valued function defined on some interval I = [a, b] containing these points. Then, there exists exactly one polynomial $P_n(x)$ of degree $\leq$ n, which interpolates f(x) at $x_o, x_1, \ldots, x_n$, that is, $P_n(x_i) = f(x_i) = f_i$ i = 0, 1, 2, …, n.

Proof: Consider the problem of determining a polynomial of degree 1 that passes through the distinct points $(x_0 y_0)$ and $(x_{1,} f_1)$. That is, we are approximating the function f by means of a first degree polynomial interpolating f(x) at $x = x_0$ and x_1.

Consider the polynomial

$$P_1(x) = \frac{(x - x_1)}{(x_0 - x_1)} f_0 + \frac{(x - x_0)}{(x_1 - x_0)} f_1$$

Then $P_1(x_0) = f_0$ and $P_1(x_1) = f_1$. Hence $P_1(x)$ has the required properties. Let

$L_0(x) = \dfrac{(x - x_1)}{(x_0 - x_1)}$ and $L_1(x) = \dfrac{(x - x_0)}{(x_1 - x_0)}$, then $P_1(x) = L_0(x)\, f_0 + L_1(x)\, f_1$.

Also we note that $L_0(x_0) = 1, L_1(x_0) = 0, L_0(x_1) = 0$ and $L_1(x_1) = 1$.

For the general case, let the required polynomial be written as

$$P_n(x) = L_0(x) f_0 + L_1(x) f_1 + \ldots + L_n(x) f_n \qquad (2.2.2)$$

$$= \sum_{i=0}^{n} L_i(x) f_i$$

Setting $x = x_j$, we get

$$f_j = P_n(x_j) = \sum_{i=0}^{n} L_i(x_j) f_i$$

Since the polynomial fits the data exactly, we must have

$$L_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

The polynomial $L_i(x)$ which are of degree $\leq$ n are called the Lagrange fundamental polynomials. It is easily verified that these polynomials are given by

$$L_i(x) = \frac{(x-x_0)(x-x_1)...(x-x_{i-1})(x-x_{i+1})...(x-x_n)}{(x_i-x_0)(x_i-x_1)...(x_i-x_{i-1})(x_i-x_{i+1}).....(x_i-x_n)}$$

$$= \prod_{\substack{j=0 \\ j\neq i}}^{n}(x-x_j) \Big/ \prod_{\substack{j=0 \\ j\neq i}}^{n}(x_i-x_j),\ i = 0, 1, 2, \ldots, n. \qquad (2.2.3)$$

Substitution of (4.3.2) in (4.3.1) gives the required **Lagrange** form of the interpolating polynomial. The uniqueness of the interpolating polynomial can be established easily with the help of following results.

Lemma 1. If $x_1, x_2, \ldots, x_k$ are distinct zeros of the polynomial P(x), then
$P(x) = (x - x_1)\ (x - x_2) \ldots (x - x_k)\ R(x)$
for some polynomial R(x)

Corollary. If $P_k(x)$ and $Q_k(x)$ are two polynomials of degree ≤ k which agree at the k +1 distinct points then $P_k(x) = Q_k(x)$ identically.

Example 1: Find the Lagrange's interpolating polynomial for the following data:

X	$\frac{1}{4}$	$\frac{1}{3}$	1
f(x)	–1	2	7

Solution:

$$L_0(x) = \frac{\left(x-\frac{1}{3}\right)(x-1)}{\left(\frac{1}{4}-\frac{1}{3}\right)\left(\frac{1}{4}-1\right)} = 16\left(x-\frac{1}{3}\right)(x-1)$$

$$L_1(x) = \frac{\left(x-\frac{1}{4}\right)(x-1)}{\left(\frac{1}{3}-\frac{1}{4}\right)\left(\frac{1}{3}-1\right)} = -18\left(x-\frac{1}{4}\right)(x-1)$$

$$L_2(x) = \frac{\left(x-\frac{1}{3}\right)\left(x-\frac{1}{4}\right)}{\left(1-\frac{1}{3}\right)\left(1-\frac{1}{4}\right)} = 2\left(x-\frac{1}{3}\right)\left(x-\frac{1}{4}\right)$$

Hence $P_2(x) = L_0(x)(-1) + L_1(x)(2) + L_2(x)(7)$

$$P_2(x) = -16\left(x-\frac{1}{3}\right)(x-1) - 36\left(x-\frac{1}{4}\right)(x-1) + 14\left(x-\frac{1}{3}\right)\left(x-\frac{1}{4}\right)$$

Example 2: If f(1) = –3, f(3) = 9, f(4) = 30 and f(6) = 132, find the Lagrange's interpolation polynomial of f(x). Also find the value of f when x = 5.

Solution: We have $x_0 = 1$, $x_1 = 3$, $x_2 = 4$, $x_3 = 6$ and
$f_0 = -3$, $f_1 = 9$, $f_2 = 30$, $f_3 = 132$.
The Lagrange's interpolating polynomial $P_3(x)$ is given by

$$P_3(x) = L_0(x) f_0 + L_1(x) f_1 + L_2(x) f_2 + L_3(x) f_3 \qquad (2.2.4)$$

Where

$$L_0(x) = \frac{(x-3)(x-4)(x-6)}{(1-3)(1-4)(1-6)} = -\frac{1}{30}(x^3 - 13x^2 + 54x - 72)$$

$$L_1(x) = \frac{(x-1)(x-4)(x-6)}{(3-1)(3-4)(3-6)} = \frac{1}{6}(x^3 - 11x^2 + 34x - 24)$$

$$L_2(x) = \frac{(x-1)(x-3)(x-6)}{(4-1)(4-3)(4-6)} = -\frac{1}{6}(x^3 - 10x^2 + 27x - 18)$$

$$L_3(x) = \frac{(x-1)(x-3)(x-4)}{(6-1)(6-3)(6-4)} = \frac{1}{30}(x^3 - 8x^n + 19x - 12)$$

Substituting these in (2.2.4) we have:

$$P_3(x) = -\frac{1}{30}(x^3 - 13x^2 + 54x - 72)(-3) + \frac{1}{6}(x^3 - 11x^2 + 34x - 24)$$
$$\times(9) - \frac{1}{6}(x^3 - 10x^2 + 27x - 18)(30)$$
$$+ \frac{1}{30}(x^3 - 8x^2 + 19x - 12(132)$$

which gives on simplification

$P_3(x) = x^3 - 3x^2 + 5x - 6.$

$$f(5) \approx P_3(5) = (5)^3 - 3(5)^2 + 5\times 5 - 6 = 125 - 75 + 25 - 6 = 69$$

You may now solve the following exercises

E1) Prove the uniqueness of the interpolating polynomial using corollary of Lemma 1.

E2) Find Lagrange's interpolating polynomial for the data. Hence obtain f(2).

x	0	1	4	5
f(x)	8	11	68	123

E3) Using the Lagrange's interpolation formula, find the value of y when x=10

x	5	6	9	11
f(x)	12	13	14	16

2.2.3 Inverse Interpolation

In inverse interpolation in a table of values of x find y = f(x), one is given a number $\bar{y}$ and wishes to find the point $\bar{x}$ so that $f(\bar{x}) = \bar{y}$, where f(x) is the tabulated function. For this, we naturally assume that the function y = f(x) has a unique inverse x = g(y) in the range of the table. The problem of inverse interpolation simply reduces to interpolation with x-row and y-row in the given table interchanged so that the interpolating points now become $y_0, y_1, \ldots, y_n$ (same as $f_0, f_1, \ldots, f_n$ i.e. $f(x_i) = f_i = y_i$) and the corresponding function values are $x_0, x_1, \ldots, x_n$ where the function is x = g(y). Since the points $y_0, y_1, \ldots, y_n$ are invariably unequally spaced, this interpolation can be done by Lagrange's form of interpolation (also by Newton's divided difference form discussed later). By Lagrange's formula

$$P_n(y) = \sum_{i=0}^{n} x_i \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(y - y_j)}{(y_i - y_j)} \text{ and}$$

$\bar{x} \approx P_n(\bar{y})$. This process is called inverse interpolation

Example 3: Find the value of x when y = 3 from the following table of values

x	4	7	10	12
y	–1	1	2	4

Solution: The Lagrange's interpolation polynomial of x is given by

$$P(y) = \frac{(y-1)(y-2)(y-4)}{(-2)(-3)(-5)}(4) + \frac{(y+1)(y-2)(y-4)}{(2)(1)(-3)}(7)$$

$$+ \frac{(y+1)(y-1)(y-4)}{(3)(1)(-2)}(10) + \frac{(y+1)(y-1)(y-2)}{(5)(3)(2)}(12)$$

So $P(3) = \frac{(2)(1)(-1)}{-(2)(3)(5)} \cdot (4) + \frac{(4)(1)(-1)}{(2)(3)} \cdot (7)$

$$+ \frac{(4)(2)(-1)}{-(3)(2)}(10) + \frac{(4)(2)(1)}{(5)(3)(2)}(12)$$

$$= \frac{4}{15} - \frac{14}{3} + \frac{40}{3} + \frac{48}{15} = \frac{182}{15} = 12.1333$$

$\therefore x(3) \approx P(3) = 12.1333.$

You may now solve the following exercises

E4) Using Lagranges interpolation formula, find the value of f(4) from the following data:

x	1	3	7	13
f(x)	2	5	12	20

E5) From the following table, find the Lagrange's interpolating polynomial, which agrees with the values of x at the given value of y. Hence find the value of x when y = 2.

x	1	19	49	101
y	1	3	4	5

E6) Using the inverse Lagrange interpolation, find the value of x when y = 3 from the following table of values:

x	36	54	72	144
y	– 2	1	2	4

2.2.4 General Error Term

The next step is to estimate/calculate a remainder term or bound for the error involved in approximating a function by an interpolating polynomial.
Let $E_n(x) = f(x) - P_n(x)$ be the error involved in approximating the function f(x) by an interpolating polynomial. We derive an expression for $E_n(x)$ in the following

theorem. We shall just indicate the proof. This result helps us in estimating a useful bound on the error as explained through an example.

Theorem 2: Let $x_0, x_1, ..., x_n$ be distinct numbers in the interval [a, b] and f has (continuous) derivatives up to order (n + 1) in the open interval (a, b). If $P_n(x)$ is the interpolating polynomial of degree ≤ n, which interpolating f(x) at the points $x_0, x_1, \ldots, x_n$, then for each $x \in [a, b]$, a number $\xi(x)$ in (a, b) exists such that

$$E_n(x) = f(x) - P_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)...(x - x_n).$$

Proof (Indication): If $x \neq x_k$ for any k = 0, 1, 2, …, n, define the function g for t in [a, b] by

$$g(t) = f(t) - P_n(t) - [f(x) - P_n(x)] \prod_{j=0}^{n} \frac{(t - x_j)}{(x - x_j)}.$$

g(t) has continuous derivatives up to (n + 1) order. Now, for k = 0, 1, 2, …, n, we have $g(x_k) = 0$ and g(x) = 0.

Thus, g has continuous derivatives up to order (n + 1) and g vanishes at the (n + 2) distinct points x, $x_0, x_1, \ldots, x_n$. By generalized Rolle's Theorem stated below, there exists $\xi \ (\xi(x))$ in (a, b) for which $g^{(n+1)}(\xi) = 0$. That is

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - (n+1)! \frac{[f(x) - P_n(x)]}{\prod_{i=0}^{n}(x - x_i)}$$

where the differentiation is with respect to t.

Simplifying above we get the error at $x = \bar{x}$

$$E_n(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = \frac{f^{(n+1)}(\xi(\bar{x}))}{(n+1)!} \prod_{i=0}^{n}(\bar{x} - x_i) \qquad (2.2.5)$$

By repeated application of Rolle's Theorem the following theorem can be proved.

Theorem 3 (Generalized Rolle's Theorem):

Let f be a real-valued function defined on [a, b] which is n times differentiable on (a, b). If f vanishes at the (n + 1) distinct prints $x_0, x_1, \ldots, x_n$, in [a, b] then a number c in (a, b) exists such that $f^{(n)}(c) = 0$.

The error formula derived above, is an important theoretical result and error formula and interpolating polynomial will be used in deriving important formula for numerical differentiation and numerical integration

It is to be noted that $\xi = \xi(\bar{x})$ depends on the point $\bar{x}$ at which error estimate is required. This error formula is of limited utility since $f^{(n+1)}(x)$ is not known (when we are given a set of data at specific nodes) and the point ξ is hardly known. But the formula can be used to obtain a bound on the error of interpolating polynomial as illustrated below:

Example 4: The following table given the values of $f(x) = e^x$, $.1 \leq x \leq 2$. If we fit an interpolating polynomial of degree four to the data given below, find the magnitude of the maximum possible error in the computed value of f(x) when x = 1.25.

x	1.2	1.3	1.4	1.5	1.6
f(x)	3.3201	3.6692	4.0552	4.4817	4.9530

Solution: From Equation, the magnitude of the error associated with the 4th degree polynomial approximation is given by

$$|E_4(x)| = |(x-x_0)(x-x_1)(x-x_2)(x-x_3)(x-x_4)\frac{f^{(5)}(\xi\xi}{5!}|$$

$$= \left|(x-1.2)(x-1.3)(x-1.4)(x-1.5)(x-1.6)\frac{f^{(5)}(\xi\xi}{5!}\right|$$

$f^{(5)}(x) = e^x$ and $\underset{1.2\le x\le 1.6}{Max}|f^{(5)}(x)| = e^{1.6} = 4.9530$

Since $f(x) = e^x$ is increasing in [1.2, 1.6]. Hence

$$|E_4(1.25)| \le (1.25-1.2)(1.25-1.3)(1.25-1.4)(1.25-1.5)\times\frac{4.9530}{120} = 0.00000135$$

When nodes are equipspaced, we shall get another bound later.

2.3 NEWTON FORM OF THE INTERPOLATING POLYNOMIAL

The Lagrange's form of the interpolating polynomial discussed in previous section has certain drawbacks. One generally calculates a linear polynomial $P_1(x)$, a quadratic polynomial $P_2(x)$ etc. by increasing the number of interpolating points, until a satisfactory approximation $P_k(x)$ to f(x) has been found. In such a situation, Lagrange form does not take any advantage of the availability of $P_{k-1}(x)$ in calculating $P_k(x)$. In Newton form, this advantage is taken care of.

Before deriving Newton's general form of interpolating polynomial, we introduce the concept of divided difference and the tabular representation of divided differences. Also the error term of the interpolating polynomial in this case is derived in terms of divided differences. Using the two different expressions for the error term we establish a relationship between nth order divided difference and the nth order derivative.

Suppose we have determined a polynomial $P_{k-1}(x)$ of degree $\le k-1$ which interpolates f(x) at the points $x_0, x_1, \ldots, x_{k-1}$. In order to make use of $P_{k-1}(x)$ in calculating $P_k(x)$ we consider the following problem. What function g(x) should be added to $P_{k-1}(x)$ to get $P_k(x)$? Let $g(x) = P_k(x) - P_{k-1}(x)$. Now, g(x) is a polynomial of degree $\le k$ and

$g(x_i) = P_k(x_i) - P_{k-1}(x_i) = f(x_i) - f(x_i) = 0$ for = 0, 1, …, k – 1.

Hence g(x) can be written as $A_k(x-x_0).....(x-x_{k-1})$
Where A_k is a constant depending on $x_0, x_1, \ldots, x_{k-1}$.

Suppose that $P_n(x)$ is the Lagrange polynomial of degree at most n that agrees with the function f at the distinct numbers $x_0, x_1, \ldots, x_n$. The divided difference of f with respect to $x_0, x_1, \ldots, x_n$ can be obtained by proving that P_n has the representation, called Newton form.

$$P_n(x) = A_0 + A_1(x-x_0) + A_2(x-x_0)(x-x_1) + ...$$
$$+A_n(x-x_0).....(x-x_{n-1}) \qquad (2.3.1)$$

for appropriate constants $A_0, A_1, \ldots..A_n$.

Evaluating $P_n(x)$ [Equation (4.4.1)] at x_o we get $A_0 = P_n(x_0) = f(x_0)$.

Similarly when $P_n(x)$ is evaluated at x_1, we get $A_1 = \frac{f(x_1) - f(x_1)}{x_1 - x_0}$. Let us introduce the notation for divided differences and define it at this stage. The divided difference of the function f, with respect to x_i, is denoted by $f[x_i]$ and is simply the evaluation f at x_i, that is, $f[x_i]=f(x_i)$. The first divided difference of f with respect to x_i and x_{i+1} is denoted by $f[x_i, x_{i+1}]$ and defined as

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i} = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}$$

The remaining divided differences of higher orders are defined inductively as follows. The kth divided differences relative to x_i, x_{i+1}, …, x_{i+k} is defined as follows:

$$f[x_i, x_{i+1}, .., x_{i+k}] = \frac{f[x_{i+1}, ... x_{i+k}] - f[x_i, ...x_{i+k-1}]}{x_{i+k} - x_i}$$

where the (k –1)st divided differences $f[x_i,…, x_{i+k–1}]$ and $f[x_{i+1},…, x_{i+k}]$ have been determined. This shows that the kth divided difference is the divided differences of (k –1)st divided differences justifying the name. It can be shown that the divided difference $f[x_0, x_{1–1}, x_k]$ is invariant under all permutations of argument $x_0, x_1,…, x_k$.

The constant $A_2, A_3, …, A_n$ can be consecutively obtained in similar manner like the evaluation of A_0 and A_1. As shown in the evaluation of A_0 and A_1 , the required constants $A_k=f[x_0, x_1,…x_k]$.

This shows that $P_n(x)$ can be constructed step by step with the addition of the next term in Equation (4.4.1), as one constructs the sequence $P_0(x)$, $P_1(x)$ with $P_k(x)$ obtained from $P_{k-1}(x)$ in the form

$$P_k(x) = P_{k-1}(x) + A_k(x - x_0).....(x - x_{k-1}) \qquad (2.3.2)$$

That is, g(x) is a polynomial of degree ≤ k having (at least) k distinct zeros $x_0,…..x_{k–1.}$ This constant A_k is called the kth divided difference of f(x) at the points $x_0, x_1, …, x_k$ and is denoted by $f[x_0, x_1…, x_k]$. This coefficient depends only on the values of f(x) at the points $x_0, x_1, …, x_k$. The following expressions can be proved for $f[x_0, x_1, …..x_k]$.

$$f[x_0, x_1.....x_k] = \sum_{i=0}^{k} \frac{f(x_i)}{(x_i - x_0)...(x_i - x_{i-1})....(x_i - x_{i+1})...(x_i - x_k)} \qquad (2.3.3)$$

and

$$f[x_0, x_1, ...x_k] = \frac{f[x_1, ...x_k] - f[x_0, ...x_{k-1}]}{x_k - x_0} \qquad (2.3.4)$$

Equation (4.4.2) can now be rewritten as

$$P_k(x) = P_{k-1}(x) + f[x_0, x_1, ..., x_k](x - x_0), ..., (x - x_{k-1}) \qquad (2.3.5)$$

Using Equation (4.4.5), Equation (4.4.1) can be rewritten as

$$P_k(x) = f[x_0] + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2]$$
$$+ ... + (x - x_0)(x - x_1)...(x - x_{n-1})f[x_0, ..., x_n] \qquad (2.3.6)$$

This can be written compactly as follows

$$P_n(x) = \sum_{i=0}^{n} f[x_0, x_1 ... x_i] \prod_{j=0}^{i-1} (x - x_j) \qquad (2.3.7)$$

This is Newton’s form of interpolating polynomial or Newton’s interpolatory divided-difference formula.

Methods for determining the explicit representation of an interpolating polynomial from tabulated data are known as divided-difference methods. These methods can be used to derive techniques for approximating the derivatives and integrals of functions, as well as for (approximating the solutions) to differential equations.

2.3.1 Table of Divided Differences

Suppose we denote, for convenience, a first order divided difference of f(x) with any two arguments by f[., .], a second order divided difference with any three arguments by f[., ., .] and so on. Then the table of divided difference can be written as follows:

Table 1

X	f[.]	f[., .]	f[., ., .]	f[[., ., ., .]	f[[., ., ., .,.]
x_0	f_0				
		$f[x_0, x_1]$			
x_1	f_1		$f[x_0, x_1, x_2]$		
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$	
x_2	f_2		$f[x_1, x_2, x_3]$		$f[x_0, x_1, x_2, x_3, x_4]$
		$f[x_2, x_3]$		$f[x_1, x_2, x_3, x_4]$	
x_3	f_3		$f[x_2, x_3, x_4]$		
		$f[x_3, x_4]$			
x_4	f4				

Example: If $f(x) = x^3$, find the value of f[*a* ,b, c].

Solution: $f[a, b] = \dfrac{f(b) - f(a)}{b - a} = \dfrac{b^3 - a^3}{b - a} = b^2 + ba + a^2$

Similarly $f[b, c] = b^2 + bc + c^2$

Hence $f[a, b, c] = \dfrac{f[b,c] - f[a,b]}{c - a}$

$$= \frac{b^2 + bc + c^2 - b^2 - ba - a^2}{c - a}$$

$$= \frac{(c - a)(c + a + b)}{(c - a)}$$

$$= a + b + c$$

You may now solve the following exercises

E7) If $f(x) = \frac{1}{x}$, show that $f[a, b, c, d] = -\frac{1}{abcd}$.

E8) Using divided difference show that the following data

x	1	2	3	5	6
f(x)	1	3	7	21	31

represents a second degree polynomial. Obtain this polynomial. Hence, find the approximate value of f(4).

Example 5: Form the following table of values, find the Newton's form of interpolating polynomial approximating f(x).

x	–1	0	3	6	7
f(x)	3	–6	39	822	1611

Also find the approximate value of the function f(x) at x = 2.

Solution: We note that the value of x, that is x_i are not equally spaced. To find the desired polynomial, we form the table of divided differences of f(x)

Table 2

x	f[.]	f[., .]	f[., ., .]	f[[., ., ., .]	f[[., ., ., .,.]
x_0 –1	3				
		–9			
x_1 0	–6		6		
		15		5	
x_2 3	39		41		1
		261		13	
x_3 6	822		132		
		789			
x_4 7	1611				

Newton's interpolating polynomial $P_4(x)$ is given by

$$P_4(x) = f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + (x - x_0)(x - x_1)(x - x_2)f[x_0, x_1, x_2, x_3] + (x - x_0)(x - x_1)(x - x_2)(x - x_3)f[x_0, x_1, x_2, x_3] \quad (2.3.8)$$

The divided differences $f[x_0]$, $f[x_0, x_1]$, $f[x_1, x_1, x_2]$, $f[x_0, x_1, x_2, x_3]$, $f[x_0, x_1, x_2, x_3]$ are those which lie along the diagonal at $f(x_0)$ as shown by the dotted line. Substituting the values of x_i and the values of the divided differences in Equation (4.4.8), we get

$$P_4(x) = 3 + (x + 1)(-9) + (x + 1)(x - 0)(6) + (x + 1)(x - 0)(x - 3)(5) + (x + 1)(x - 0)(x - 3)(x - 6)(1)$$

This on simplification gives

$$P_4(x) = x^4 - 3x^3 + 5x^2 - 6$$

$$\therefore f(x) \approx P_4(x)$$

$$f(2) \approx P_4(2) = 16 - 24 + 20 - 6 = 6$$

You may now solve the following exercises.

E9) From the table of values given below, obtain the value of y when x = 1 using
(a) divided difference interpolation formula.
(b) Lagrange's interpolation formula

x	0	2	3	4
f(x)	– 4	6	26	64

We now give another expression for the error term, that is, the error committed in approximating f(x) by $P_n(x)$.

Let $P_n(x)$ be the Newton form of interpolating polynomial of degree ≤ n which interpolates f(x) at $x_0, \ldots, x_n$. The Interpolating error $E_n(x)$ of $P_n(x)$ is given by

$$E_n(x) = f(x) - P_n(x). \quad (2.3.9)$$

Let $\bar{x}$ be any point different from $x_0, x_1, \ldots x_n$. If $P_{n+1}(x)$ is the Newton form of interpolating polynomial which interpolates f(x) at $x_0, x_1, \ldots x_n$ and $\bar{x}$, then $P_{n+1}(\bar{x}) = f(\bar{x})$. Then by Equation (4.4.5) we have

$$P_{n+1}(x) = P_n(x) + f[x_0, x_1, \ldots, x_n, \bar{x}] \prod_{j=0}^{n} (\bar{x} - x_j)$$

Putting $x = \bar{x}$, in the above, we have

$$f(\bar{x}) = P_{n+1}(\bar{x}) = P_n(\bar{x}) + f[x_0, x_1, \ldots x_n, \bar{x}] \prod_{j=0}^{n} (\bar{x} - x_j)$$

That is, $E_n(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = f[x_0, x_1, \ldots, x_n, \bar{x}] \prod_{j=0}^{n} (\bar{x} - x_j)$

This shows that the error is like the next term in the Newton form.

Comparing the two expressions derived for $E_n(x)$, we establish a relationship between divided differences and the derivatives of the function as follows:

$$E_n(\bar{x}) = \frac{f^{n+1}(\xi(\bar{x}))}{(n+1)} \prod_{j=0}^{n} (\bar{x} - x_i)$$

$$= f[x_0, x_1 \ldots\ldots x_n, \bar{x}] \prod_{j=0}^{n} (\bar{x} - x_i).$$

Comparing, we have

$$f[x_0, x_1, \ldots, x_{n+1}] = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

(Considering $\bar{x} = x_{n+1}$). Further, it can be shown that $\xi \in (\min x_i, \max x_i)$.
We state these results in the following theorem.

Theorem 4: Let f(x) be a real-valued function, defined on [a, b] and n time differentiable in (a, b). If $x_0, x_1, \ldots x_n$ are n+1 distinct points in [a, b], then there exists $\xi \in (a, b)$ such that

$$f[x_0, \ldots x_n] = \frac{f^{(n)}(\xi)}{n!}$$

Corollary 1: If $f(x) = x^n$, then $f[x_0, x_1, \ldots, x_n] = \frac{n!}{n!} = 1$

Corollary 2: If $f(x) = x^k$, $k < n$, then $f[x_0, \ldots, x_k] = 0$
(Since the nth derivative of x^k, $k < n$, is zero).

Example 6: If $f(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$, then find $f[x_0, x_1, \ldots, x_n]$.

Solution: By corollaries 1 and 2, we have $f[x_0, x_1, \ldots, x_n] = a_n \frac{n_1}{n_1} + 0 = a_n$

Let us consider the error formula

$$E_n(x) = f(x) - P_n(x) = \frac{f^{n+1}(\xi(x))}{(n+1)!} \prod_{j=0}^{n} (x - x_i)$$

We want to obtain a bound on the error. For this we have

$$|E_n(x)| = |f(x) - P_n(x)| \le \frac{\max_{t \in I} |f^{(n+1)}(t)|}{(n+1)!} \max_{t \in I} |\psi_n(t)| \qquad (2.3.10)$$

where [a, b] = I, $\psi_n(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$.

Example 7: Find a bound for the error in linear interpolation.

Solution: For the case n = 1, we have linear interpolation. If $x \in [x_{i-1}, x_i]$ then we approximate f(x) by $P_1(x)$ which interpolates at x_{i-1}, x_i. From Equation (4.4.10). We have

$$|E_1(x)| \le \frac{1}{2} \max_{x \in I} |f''(x)| \max_{x \in I} |\psi_1(x)|$$

where $\psi_1(x) = (x - x_{i-1})(x - x_i)$.

Now $\frac{d\psi_1}{dx} = x - x_{i-1} + x - x_i = 0 \Rightarrow x = (x_{i-1} + x_i)/2$.

Hence, the maximum value of $|(x - x_{i-1})(x - x_i)|$ occurs at $x = x^* = (x_{i-1} + x_i)/2$.

$$\therefore \max_{x_{i-1} \le t \le x_i} |\psi_1(x)| = \frac{(x_i - x_{i-1})^2}{4} = \frac{h^2}{4}.$$

Thus $|E_1(x)| \le \frac{(x_i - x_{i-1})^2}{4} \cdot \frac{1}{2} \max_{x \in I} |F''(x)| = \frac{h^2}{8} \cdot M$

Where $h = x_i - x_{i-1}$ and $|f''(x)| \le M$ on $I=[x_{i-1}, x_i]$.

Example 8: Determine the largest step-size h that can be used in the tabulation of the function $f(x) = (1 + x^2)^2$ on [0, 1] so that the error in the linear interpolation is less than 5×10^{-5}.

Solution: We have to determine the spacing h in a table of equally spaced values of the function of $f(x) = (1 + x^2)^2$ between 0 and 1. By assumption, the table will contain $f(x_i)$ with $x_i = 1 + ih$, i = 0, …, N, where $N = \frac{1-0}{h} = \frac{1}{h}$.

If $\bar{x} \in [x_{i-1}, x_i]$, we approximate $f(\bar{x})$ by $P_1(\bar{x})$, where $P_1(x)$ is the linear polynomial which interpolates f(x) at x_{i-1} and x_i. By Equation (4.4.10)

$$|E_1(\bar{x})| \le \frac{h^2}{8} \cdot M \le \frac{h^2}{8} \cdot \max_{x \in [0,1]} |f''(x)|$$

When $M = \max_{x \in [x_{i-1}, x_i]} |f''(x)|$.

$$|E_1(\bar{x})| \le \frac{h^2}{8} \cdot M' \text{ where } M' = \max_{0 \le x \le 1} |f''(x)|$$

When $\bar{x} \in (0, 1)$.

Since $f''(x) = 12x^2 + 4$ and it is an increasing function on [0, 1], $\max_{x \in [0,1]} |f''(x)| = 16$.

Thus $|E_1(\bar{x})| \le \frac{h^2}{8} \cdot 16 = 2h^2$

We have $2h^2 < 5 \times 10^{-5}$ or $h \le .005$.

That is $N = \frac{1}{.005} = \frac{1000}{5} = 200$.

You may now solve the following exercises.

E10) Determine the spacing h in a table of equally spaced points of the function $f(x) = x\sqrt{x}$ between 1 and 2, so that the error in the linear interpolation is less than 5×10^{-6}.

2.4 INTERPOLATION AT EQUALLY SPACED POINTS

Suppose the value of f(x) at (n + 1) equally spaced values of x are known or given, that is, (x_i, y_i), i = 0,…, n are known where $x_i - x_{i-1} = h$ (fixed), i = 1, 2, …, x and y = f(x). Suppose we are required to approximate value of f(x) or its derivative f'(x) for some values of x in the interval of interest. The methods for solving such problems are based on the concept of finite differences. However, our discussion will be confined to obtain Newton's forward and backward difference forms only. Each is suitable for use under specific situation.

2.4.1 Differences – Forward and Backward Differences

Suppose that we are given a table of values (x_i, y_i) i = 0,…, N where $y_i = f(x_i) = f_i$. Let the nodal points be equidistant. That is $x_i = x_0 + ih$, i = 0, 1, …, N with $N = (b - a)/h$ (I = [a, b]).

For simplicity we introduce a linear change of variables $s = s(x) = \dfrac{x - x_0}{h}$, so that x = x(s) = x_0 + sh and introduce the notation.

$f(x)=f(x_0 + sh)=f_s$.

The linear change of variables transforms polynomials of degree n in x into polynomial of degree n in s. For equally spaced nodes, we shall deal with two types of differences, namely forward and backward and discuss their representation in the form of a table. We shall also derive/give the relationship of these differences with divided differences and their interrelationship.

Forward Differences

We denote the forward differences of f(x) of ith order at $x = x_0 + sh$ by $\Delta^i f_s$ and define it as follows:

$$\Delta^i f_s = \begin{cases} f_s, & i=0 \\ \Delta(\Delta^{i-1} f_s) = \Delta^{i-1} f_{s+1} - \Delta^{i-1} f_s, & i > 0 \end{cases} \qquad (2.4.1)$$

where Δ denotes forward difference operator.

When s = k, that is $x = s_k$, we have

for i = 1 $\quad \Delta f_k = f_{k+1} - f_k$

for i =2 $\quad \Delta^2 f_k = \Delta f_{k+1} - \Delta f_k = f_{k+2} - f_{k+1} - [f_{k+1} - f_k]$

$= f_{k+2} - 2f_{k+1} + f_k$

Similarly, you can obtain

$$\Delta^3 f_k = f^{k+3} - 3 f_{k+2} + 3f_{k+1} - f_k.$$

We recall the binomial theorem

$$(a+b)^s = \sum_{j=0}^{s} \binom{s}{j} a^j b^{s-j}. \qquad (2.4.2)$$

where $\binom{s}{j} = C(s, j)$

and s is a real and non-negative integer.

The shift operator E is defined as

$$E f_i = f_{i+I} \qquad (2.4.3)$$

In general $Ef(x) = f(x + h)$. We have $E^s f_i = f_{i+s}$

For example $E^4 f_i = f_{i+4}$. $E^{1/2} f_i = f_{i+1/2}$ and $E^{-1/2} f_i = f_{i-\frac{1}{2}}$

Now $\Delta f_i = f_{i+1} - f_i = Ef_i - f_i = (E-1) f_i$

Hence the shift and forward difference operators are related by
$\Delta = E - 1$ or $E = 1 + \Delta$.
Operating s times, we get

$$\Delta^s = (E-1)^s = \sum_{j=0}^{s} \binom{s}{j} E^j (-1)^{s-j} \qquad (2.4.4)$$

by Equation (4.5.2) (4.5.3), we get

$$\Delta^s f_i = \sum_{j=0}^{s} \binom{s}{j} (-1)^{s-j} \binom{s}{j} f_{j+1}$$

We now give in the following table, the forward differences of various orders using 5 values.

Table 3 : Forward Difference Table

x	f(x)	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
x_0	f_0				
		Δf_0			
x_1	f_1		$\Delta^2 f_0$		
		Δf_1		$\Delta^3 f_0$	
x_2	f_2		$\Delta^2 f_1$		$\Delta^4 f_0$
		Δf_2		$\Delta^3 f_1$	
x_3	f_3		$\Delta^2 f_2$		
		Δf_3			
x_4	f_4				

Note that the forward difference $\Delta^k f_0$ lie on a straight line slopping down ward to the right.

Now we give Lemma 1 the relationship between the forward and divided differences, which can be proved by induction. This relation is utilized to derive the Newton's forward-difference formula which interpolates f(x) at
$x_k + ih$, i =0, 1, …, n.

Lemma 1: For all in ≥ 0

$$f[x_k, x_{k+1}, x_{k+i}] = \frac{1}{i! h^i} \Delta^i f_k$$

This has an easy corollary.

Corollary: If $P_n(x)$ is a polynomial of degree n with leading co-efficients a_n, and x_0 is an arbitrary point, then
$\Delta^n P_n(x_0) = a_n\, n!\, h^n$ and
$\Delta^{n+1} P_n(x_0) = 0$, i.e., all higher differences are zero.

Backward Differences

The backward differences of f(x) of ith order at $x_s = x_0 + sh$ are denoted by $\nabla^i f_s$. They are defined as follows:

$$\nabla^i f_s = \nabla^i f_s = \left\{ \nabla^{i-1}(\nabla f_s) = \nabla^{i-1}[f_s - f_{s-1}], i \geq 1 \right\} = \nabla^{t-1} f_s - \nabla^{t-1} f_{s-1} \quad (2.5.5)$$

where ∇ denotes backward difference operator. When s = k, that is $x = x_0 + kh = x_k$, we have for

i = 1 $\quad \nabla f_k = f_k - f_{k-1}$

i = 2 $\quad \nabla^2 f_k = \nabla(f_k - f_{k-1}) = \nabla f_k - \nabla f_{k-1} = f_k - f_{k-1} + f_{k-2}$

Similarly for

i = 3 $\quad \nabla^3 f_k = f_k - 3f_{k-1} + 3f_{k-2} - f_{k-3}$

The relation between the backward difference operator ∇ and the shift operator E is given by

$$\nabla = 1 - E^{-1} \qquad E = (I - \nabla)^{-1}$$

Also $\Delta = E - 1 = E(I - E^{-1}) = E\nabla$

Since $\nabla f_k = f_k - f_{k-1} = f_k - E^{-1} f_k = (1 - E^{-1}) f_k$

Operating s time, we get

$$\nabla^s f_k = (1 - E^{-1})^s f_k = \sum_{j=0}^{s} (-1)^j \binom{s}{j} f_{k-j}$$

We can extend the binomial coefficient notation to include negative numbers, by letting

$$\binom{-s}{i} = \frac{-s(-s-1)(-s-2)...(-s-i+1)}{i!} = (-1)^i \frac{s(s+1)...(s+i-1)}{i!}$$

The backward differences of various orders with 5 nodes are given in the following table:

Table 4: Backward Difference Table

x	f(x)	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$
x_0	f_0				
		∇f_1			
x_1	f_1		$\nabla^2 f_2$		
		∇f_2		$\nabla^3 f_3$	
x_2	f_2		$\nabla^2 f_3$		$\nabla^4 f_4$
		∇f_3		$\nabla^3 f_4$	
x_3	f_3		$\nabla^2 f_4$		
		∇f_4			
x_4	f_4				

Note that the backward difference $\nabla^k f_4$ lie on a straight line slopping upward to the right. Also note that $\Delta f_k = \nabla f_{k+1} = f_{k+1} - f_k$.

2.4.2 Newton's Forward-Difference and Backward-Difference Formulas

The Newton's form of interpolating polynomial interpolating at x_k, x_{k+1} ., x_{k+n} is

$$P_n(x) = \sum_{i=0}^{n} (x - x_k)(x - x_{k+1})....(x - x_{k+i-1}) f[x_{k+1}, x_{k+2}, ..., x_{k+1}] \quad (2.4.6)$$

In this we make use of the following:

$f[x_k, ..., x_{k+n}] = \frac{1}{n! h^n} \Delta^n f_k$ and get

$$P_n(x) = \sum_{i=0}^{n} (x-x_k)(x-x_{k+1}) \ldots (x-x_{k+i-1}) \frac{1}{i!h^n} \Delta^i f_k \qquad (2.4.7)$$

Setting k = 0, we have

Here $x_s = x_0 + sh$ may be introduced, where

$$s = \frac{x_s - x_0}{h}$$

Also f(x) = can be derived straight from $E^s f_0 = (1 + \Delta)^s f_0$

$$P_n(x) = f_0 + \frac{(x-x_0)}{1!}\frac{\Delta f_0}{h} + \frac{(x-x_0)(x-x_1)}{2!}\frac{\Delta^2 f_0}{h^2} + \ldots \frac{(x-x_0)\ldots(x-x_{n-1})}{n!}\frac{\Delta^n f_0}{h^n} \qquad (2.4.8)$$

Also we have $x - x_{k+j} = x_0 + sh - [x_0 + (k+j)h] = (s-k-j)h$

Substituting this in Equation (4.5.8) and simplifying we get

$$\begin{aligned} P_n(x) = p(x_0 + sh) &= \sum_{i=0}^{n} (s-k)(s-k-1)\ldots(s-k-i+1)\Delta^i f_k \\ &= \sum_{i=0}^{n} \Delta^i f_k \binom{s-k}{i} \\ &= f_k + (s-k)\Delta)_k + \frac{(s-k)(s-k-1)}{2!}\Delta^2 f_k + \ldots \\ &\quad + \frac{(s-k)\ldots(s-n-1)}{n!}\Delta^n f_k \end{aligned} \qquad (2.4.9)$$

of degree ≤ n.

Setting k = 0, in Equation (4.5.9) we get the formula

$$P_n(x_0 + sh) = \sum_{i=0}^{n} \Delta^i f_0 \binom{s}{i}$$

This form Equation (4.5.1)..(4.5.9)…. is called the Newton's forward-difference formula.

The error term is now given by

$$E_n(x) = \binom{s}{n+1} h^{n+1} f^{n+1}(\xi\xi$$

For deriving the Newton's Backward-Difference formula, we reorder the interpolating nodes as $x_n, x_{n-1}, \ldots . x_0$ and apply the Newton's divided difference form. This gives

$$\begin{aligned} P_n(x) &= f[x_n] + (x-x_n) f[x_{n-1}, x_n] + (x-x_n)(x-x_{n-1}) f[x_{n-2}, x_{n-1}, x_n] \\ &\quad + (x-x_n) \ldots . (x-x_0) f[x_{n \ldots \ldots . x_0}] \end{aligned} \qquad (2.4.10)$$

Set x = x_n + sh, then

$$x - x_j = x_n + sh - [x_n - (n-i)h] = (s+n-j)h$$

$$x - x_{n-j} = x_n + sh - [x_n - (n-n+j)h] = (s+j)h$$

and

$$(x-x_n)(x-x_{n-1})\ldots(x-x_{n-i+1}) = s(s+1)\ldots(s+i-1)h^x$$

Also we have $f[x_{n-k}, \ldots, x_n] = \frac{1}{k!h^k}\Delta^k(x_n) = \frac{1}{k!h^k}\Delta^k f(x_n)$

Substituting those in Equation (2.4.10) and simplifying we get the following two expressions

$$\begin{aligned} P_n(x) &= \sum_{i=0}^{n} \frac{1}{i!h^i}(x-x_n)(x-x_{n-i})\ldots(x-x_{n-i+1})\Delta^i f_n \\ &= f_n + \frac{(x-x_n)}{1!}\frac{\Delta f_n}{h} + \frac{(x-x_n)(x-x_{n-i})}{2!}\frac{\Delta^2 f_n}{h^2} + \ldots \end{aligned}$$

$$+\frac{(x-x_n)(x-x_{n-i})\ldots(x-x_n)}{n!}\frac{\Delta^n f_n}{h^n} \qquad (2.4.11)$$

$$=f_n+s\Delta\Delta_n+\frac{s(s+1)}{2!}\Delta^2 f_n+\ldots+\frac{s(s+1)\ldots(s+n-1)}{n!}\Delta^n f_n$$

Since $\binom{-s}{k}=\frac{(-1)^k s(s+1)\ldots(s+k-1)}{k!}$, we have

$$P_n(x)=f(x_n)+(-1)\binom{-s}{1}\nabla f(x_n)+(-1)^2\binom{-s}{2}\nabla^2 f(x_n)+\ldots+(-1)^n\binom{-s}{n}\nabla^n f(x_n) \qquad (2.4.12)$$

or

$$P_n(x)=\sum_{k=0}^{n}(-1)^k\binom{-s}{k}\nabla^k f(x_n)$$

This is called Newton's backward-difference form.

In this case, error is given by

$$E_n(x)=(-1)^{n+1}\frac{s(s+1)\ldots(s+n)}{(n+1)!}h^{n+1}f^{n+1}(\xi\xi)$$

The forward-difference formula (4.5.7) is suitable for approximating the value of the function at x that lies towards the beginning of the table and the backward-difference form is suitable for approximating the value of the function at x that lies towards the end of the table.

Example 9: Find the Newton's forward-difference interpolating polynomial which agrees with the table of values given below. Hence obtain the value of f(x) at x = 1.5.

x	1	2	3	4	5	6
f(x)	10	19	40	79	142	235

Solution:

Table 5: Forward Differences

x	f(x)	Δf	Δ^2f	Δ^3f		
1	10					
		9				
2	19		12			
		21		6		
3	40		18		0	
		39		6		0
4	79		24		0	
		63		6		
5	142		30			
		93				
6	285					

The Newton's forward-differences interpolation polynomial is

$$f(x)\approx f_0+(x-1)\Delta)_0+\frac{(x-1)(x-2)}{2!}\Delta^2 f_0+\frac{(x-1)(x-2)(x-3)}{3!}\Delta^3 f_0$$

$$=10+(x-1)(9)+\frac{(x-1)(x-2)}{2}(12)+\frac{(x-1)(x-2)(x-3)}{6}(6)$$

On simplification we get

$f(x) \approx x^3 + 2x + 7$

$\therefore f(1.5) = (1.5)^3 + 2(1.5) + 7 = 13.375$

Example 10: Find Newton's backward difference form of interpolating polynomial for the data

x	4	6	8	10
f(x)	19	40	79	142

Hence interpolate f(9).

Solution: We have

Table 6: Backward Difference

x	f(x)	∇f	$\nabla^2 f$	$\nabla^3 f$
4	19			
		21		
6	40		18	
		39		6
8	79		24	
		63		
10	142			

$$P_n(x) = 142 + (x-10)\frac{63}{2} + \frac{(x-10)(x-8)}{2!}\cdot\frac{24}{4} + \frac{(x-10)(x-8)(x-6)}{3!}\cdot\frac{6}{8}$$

$$f(9) \approx P_n(9) = 142 - \frac{63}{2} - 3 - \frac{3}{8} = 107.125$$

You may now solve the following exercises.

E11) Find the Newton's backward differences interpolating polynomial for the data of Example 9.

E12) Using forward differences, show that the following data represents a third degree polynomial.

x	– 3	– 2	– 1	0	1	2	3
f(x)	– 29	– 9	– 1	1	3	11	31

Find the polynomial and obtain the value of f(0.5).

E13) Using forward Differences, show that the following data:

x	– 1	0	1	2	3	4
f(x)	6	1	0	3	10	21

represents a second degree polynomial. Find this polynomial and an approximate value of f(2.5)

E14) Estimate the value of f(1.45) from the data given below

x	1.1	1.2	1.3	1.4	1.5
f(x)	1.3357	1.5095	1.6984	1.9043	2.1293

E15) Evaluate the differences

(i) $\nabla^3[a_2x^2 + a_1x + a_0]$

(ii) $\nabla^3[a_3x^3 + a_2x^2 + a_1x + a_0]$

(iii) $\Delta^3[a_3x^3 + a_2x^2 + a_1x + a_0]$

E16) Show that the nth order divided differences of f(x) = $\frac{1}{x}$ is

$(-1)^n /(x_0.x_1...x_n)$.

E17) A table of values is to be constructed for the function f(x) given by $f(x) = x^4 + 1$ in the interval [3, 4] with equal step-length. Determine the spacing h such that linear interpolation gives results with accuracy 1×10^{-4}.

E18) Find the Newton's divided difference form of interpolating polynomial for the data.

x	– 4	–1	0	2	5
f(x)	1245	33	5	9	1335

Also approximate f(1) from the polynomial.

E19) Construct Newton's forward difference table for the data

x	3	5	7	9
f(x)	6	24	38	108

Hence approximate f(4) from Newton's forward difference interpolating polynomial.

E20) If $f(x) = ax^2 + bx + c$ (a, b, c, $\in$, R), then show that f[1, 2, 3] = a.

2.5 SUMMARY

In the first section apart from deriving the Lagrange's form of interpolating polynomial for a given data, it has been shown that the interpolating polynomial for a given data is unique. We have also seen how the Lagrange's interpolation formula can be applied with y as the independent variable and x as the dependent variable so that the value of x corresponding to a given value of y can be calculated approximately when some conditions are satisfied. Finally, we have derived the general error formula and its use has been illustrated to judge the accuracy of our calculation. In the next section, we have derived a form of interpolating polynomial called Newton's general form (divided difference form) which has some advantages over the Lagrange's form discussed in section 1. We have introduced the concept of divided differences and discussed some of its properties before deriving Newton's general form. The error term also has been derived and we have established a relationship between the divided difference and the derivative of the function f(x) using the two different expressions of the error terms. In section 3, we have derived interpolation formulas for data with **equally spaced** values of the argument. The application of the formulas derived in this section is easier compared to the application of the formulas derived in first and second sections.

The mathematical formulas derived in the unit are listed below for your easy reference.

1. Lagrange's Form

$$P_n(x) = \sum_{i=0}^{n} f(x_i)\, L_i(x) \quad \text{where} \quad L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

2. Inverse Interpolation

$$P_n(y) = \sum_{i=0}^{n} x_i \left[\prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(y - y_j)}{(y_i - y_j)} \right.$$

3. Interpolation Error

$$E_n(x) = f(x) - P_n(x) = \frac{f^{(n+1)}(\xi\xi(x)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)$$

4. Divided Difference

$$f[x_0, x_1, \ldots x_j] = \frac{f[x_1, \ldots . x_j] - f[x_0, \ldots . x_{j-1}]}{x_j - x_0}$$

5. Newton's Form

$$P_n(x) = \sum_{i=0}^{n} f[x_0, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j)$$

6. Interpolation Error (In terms of divided difference)

$$E_n(x) = f[x_0, x_1, \ldots . x_n, x] \prod_{j=0}^{n} (x - x_j)$$

7. $f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi\xi}{n!}, \xi \in (\min x_i, \max x_i)$

8. Newton's Forward Difference Formula

$$P_n(x) = P_n(x_0 + sh) = \sum_{i=0}^{n} \binom{s}{i} \Delta^i f_0$$

$$= f_0 + s\Delta\Delta_0 + \frac{s(s-1)}{2!} \Delta^2 f_0 + \ldots + \frac{s(s-1) \ldots . (s-n+1)}{n!} \Delta^n f_0$$

When $s = (x - x_0)/h$ and

$$E_n(x) = \binom{s}{n+1} h^{n+1} f^{(n+1)}(\xi\xi$$

9. Newton's Backward Difference Formula:

$$P_n(x) = P_n(x_n + sh) = \sum_{k=0}^{n} (-1)^k \binom{-s}{k} \nabla^k f_n$$

where $s = (x - x_n)/h$ and

$$E_n(x) = (-1)^{n+1} \frac{s(s+1)...(s+n)}{(n+1)!} h^{n+1} f^{(n+1)}(\xi)$$

10. $f[x_k, ..., x_{k+n}] = \frac{1}{n! h^n} \Delta^n f_k$ and

$$f[x_{n-k}, ..., x_n] = \frac{1}{k! h^k} \nabla^k f(xn).$$

2.6 SOLUTIONS/ANSWERS

E1) We show that if P(x) and Q(x) are two polynomials of degree $\leq$ k which agree at the k + 1 distinct points $x_0, x_1, \ldots x_k$, then P(x) = Q(x) identically. Let ψ (x) = P(x) = Q(x). Then ψ (x) is a polynomial of degree $\leq$ k, and by Lemmal, it can be written as ψ (x) = (x – x_0) (x – x_1)… (x – x_k) R(x) with R(x) some polynomial. Let $R(x) = c_m x^m + = c_{m-1} x^{m-1} + \ldots + c_1 x + c_0$ such that

$c_m \neq 0$. Then k $\geq$ degree of ψ = k + 1 + m which is absurd. Hence ψ (x) = 0 identically, so P(x) = Q(x).

E2) $x^3 - x^2 + 3x + 8,$ 18

E3) 14.6667

E4) 6.6875

E5) Let x = g(y). The Lagrange's interpolating polynomial P(y) of g(y) is given by

$$P(y) = -\frac{1}{24}(y^3 - 12y^2 + 47y - 60) + \frac{19}{4}(y^3 - 10y^2 + 29y - 20)$$
$$-\frac{49}{3}(y^3 - 9y^2 + 23y - 15) + \frac{101}{8}(y^3 - 8y^2 + 19y - 12)$$

which, on simplification gives

$P(y) = y^3 - y^2 + 1$ when y = 2, x $\approx P(2) = 5$.

E6) x = g(y)

$$P(y) = \frac{(y-1)(y-2)(y-4)}{(-3)(-4)(-6)}.(36) + \frac{(y+2)(y-2)(y-4)}{(3)(-1)(-3)}.(54)$$
$$+\frac{(y+2)(y-1)(y-4)}{(4)(1)(-2)}.(72) + \frac{(y+2)(y-1)(y-2)}{(6)(3)(2)}.(144)$$

$$x \approx P(3) = \frac{(2)(1)(-1)}{-72} \times (36) + \frac{(5)(1)(-1)}{9} \times (54)$$

$$+\frac{(5)(2)(1)}{8} \times (72) + \frac{(5)(2)(1)}{36} \times (144)$$

$$= 1 - 30 + 90 + 40 = 101$$

E7) $f[a,b] = \dfrac{\frac{1}{b} - \frac{1}{a}}{b-a} = -\dfrac{1}{ab}$

$f[b,c] = -\dfrac{1}{bc}$ and $f[c,d] = -\dfrac{1}{cd}$

$$f[a,b,c] = \frac{1}{abc}, f[b,c,d] = \frac{1}{bcd}$$

$$f[a,b,c,d] = \frac{f[b,c,d,] - f[a,b,c]}{d-a} = -\frac{1}{abcd}$$

E8) Divided Difference Table

x	f[x]	f[., .]	f[., ., .]		
1	1				
		2			
2	3		1		
		4		0	
3	7		1		0
		7		0	
5	21		1		
		10			
6	31				

Since third and higher order divided differences are zeros, given data represents a second degree polynomial

$$P_2(x) = f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2]$$

$= 1 + (x-1).2 + (x-1)(x-2).1 = x^2 - x + 1 \quad f(4) \approx P_2(4) = 13$

$f(4) \approx P_2(4) = 13$

E9) Lagrange's interpolation polynomial

$$P(x) = \frac{1}{6}(x^3 - 9x^2 + 26x - 24) + \frac{3}{2}(x^3 - 7x^2 + 12x)$$

$$-\frac{26}{3}(x^3 - 6x^2 + 8x) + 8(x^3 - 5x^2 + 6x)$$

$= x^3 + x - 4$ on simplification.

Divided difference interpolation polynomial:

x	f[x]	f[., .]	f[., ., .]	f[., ., .,.]
0	-4			
		5		
2	6		5	
		20		1
3	26		9	
		38		
4	64			

$P(x) = -4 + (x-0)5 + x(x-2)(5) + x(x-2)(x-3)(1)$

$= x^3 + x - 4$

$f(1) \approx p(1) = -2$.

E10) $|E_1(\bar{x})| \le \frac{h^2}{8} M$ where $M = \max_{1 \le x \le 2} = |f''(x)|$

and $\bar{x} \varepsilon (1,2).\ f^1(x) = \frac{3}{2} x^{1/2}, f''(x) = \frac{3}{4} x^{-1/2}$.

$\max_{1 \le x \le 2} |f''(x)| = \frac{3}{4}$

$|E_1(\bar{x})| \le \frac{h^2}{8} \cdot \frac{3}{4} \le 5 \cdot 10^{-6} = \frac{1}{2} \cdot 10^{-5}$

i.e. $h^2 \le \frac{32}{2 \times 3} \cdot 10^{-5} = \frac{16}{3} 10^{-5} \quad \Rightarrow h = 0.0073$

E11) **Backward Difference Table**

	x	f(x)	∇f	$\nabla^2 f$	$\nabla^3 f$
x_0	1	10			
			9		
x_1	2	19		12	
			21		6
x_2	3	40		18	0
			39		6 … 0
x_3	4	79		24	0
			63		6

x_4	5	142		30	
			93		
x_5	6	235			

$$P(x) = f_5 + (x - x_5)\nabla f_5 + \frac{(x - x_5)(x - x_4)}{2!}\nabla^2 f_5$$

$$+ \frac{(x - x_5)(x - x_4)(x - x_3)}{3!}\nabla^3 f_5 = 235 + 93(x - 6) + 15(x - 6)(x - 5) + (x - 4)(x - 5)(x - 6) = x^3 + 2x + 7$$

E12) **Forward Difference Table**

x	f(x)	Δf	Δ^2f	Δ^3f	Δ^4f
–3	–29				
		20			
–2	– 9		–12		
		8		6	
–1	– 1		– 6		0
		2		6	
0	1		0		0
		2		6	
1	3		6		0
		8		6	
2	11		12		
		20			
3	31				

$$f(x) \approx P_3(x) = f(x_0) + \frac{(x - x_0)}{h}\Delta f_0 + \frac{(x - x_0)(x - x_1)}{2!h^2}\Delta^2 f_0$$

$$+ \frac{(x - x_0)(x - x_1)(x - x_2)}{3!h^3}\Delta^3 f_0$$

$$= x^3 + x + 1 f(0.5) \approx P_3(0.5) = 1.625$$

E13) **Forward Difference Table**

x	f(x)	Δf	Δ^2f	. . .
–1	6			
		–5		
0	1		4	
		–1		0
1	0		4	
		3		0
2	3		4	
		7		0
3	10		4	
		11		
4	21			

Since third and higher order differences are zeros, f(x) represents a second order polynomial

$$f(x) \approx P_2(x) = f_0 + \frac{(x-x_0)}{h}\Delta f_0 + \frac{(x-x_0)(x-x_1)}{2!h^2}\Delta^2 f_0$$

$$= 6 + (x+1)(-5) + \frac{(x+1)(x-0)}{2}(4) = 2x^2 - 3x + 1$$

$f(2.5) \approx 6.$

E14) Backward Difference Table

x	f(x)	∇ f	∇^2f	∇^3f	∇^3f
1.1	1.3357				
		0.1738			
1.2	1.5095		0.0151		
		0.1889		0.0019	
1.3	1.6984		0.0170		0.0002
		0.2059		0.0021	
1.4	1.9043		0.0910		
		0.2050			
1.5	2.1293				

Here $x_n = 1.5, x = 1.45, h = 0.1 \therefore s = \frac{x - x_n}{h} = \frac{1.45 - 1.5}{0.1} = -0.5$

$$f(x) = f_n + s\nabla f_n + \frac{s(s+1)}{2!}\nabla^2 f_n + \frac{s(s+1)(s+2)}{3!}\nabla^3 f_n$$

$$+ \frac{s(s+1)(s+2)(s+3)\Delta^4 f_n}{4!}$$

$$= 2.1293 - 0.1125 - 0.00239 - 0.00013 - 0.0000075$$

$$= 2.01427 \approx 2.0143$$

E15) i) 0

ii) $a_3\, 3!\, h^2$ (Recall that $f(x_{0,\ldots\ldots}x_i] = \frac{f^i(\xi)}{i!}$

iii) $a_3\, 3! h^3$ and consider x fixed)

E16) Prove this by induction.

E17) $|E_1(x)| \le \frac{h^2}{8} M_2$ where $M_2 = \underset{3 \le x \le 4}{Max} |f''(x)|$

We have $f'(x) = 4x^3$, $f''(x) = 12x^2$ $M_2 = 12 \times 16 = 192$

We have $\frac{h^2}{8} \times 192 \le 10^{-4} \Rightarrow h \le \frac{1}{\sqrt{24}} 10^{-2} \approx 0.002$

E18) **Divided Difference Table**

x	f[.]	f[., .]	f[., ., .]	f[[., ., ., .]	f[[., ., ., .,.]
–4	1245				
		–404			
– 1	33		94		
		– 28		– 14	

0	5		10		3
		2		13	
2	9		88		
		442			
5	1355				

$P_4(x) = 1245 - 404(x+4) + 94(x+4)(x+1) - 14(x+4)(x+1)x + 3(x+4)(x+1)x(x-2)$

$f(1) \approx P_4(1) = -5$

E19) **Forward Difference Table**

x	f(x)	Δf	Δ^2f	Δ^3f
3	6			
		18		
5	24		–4	
		14		60
7	38		56	
		70		
9	108			

$$s = \frac{x - x_0}{h} = \frac{4-3}{2} = \frac{1}{2}$$

$$f(x_0 + sh) \approx P_4(x_0 + sh) = f_0 + \frac{1}{2}\Delta f_0 + \frac{1}{8}\Delta^2 f_0 + \frac{1}{16}\Delta^3 f_0$$

$$\text{i.e., } f(4) = 6 + \frac{1}{2}\cdot 18 - \frac{1}{8}\cdot -4 + \frac{1}{16}.60$$

$$= 6 + 9 + \frac{1}{2} + \frac{15}{4} = 19.25$$

E20) We have $f[x_0, ..., x_n] = \frac{1}{n!} f^{(n)}(\xi\xi) \in (\min x_i, \max x_i)$

$$f[1,2,3] = \frac{1}{2!} f''(\xi\xi = \frac{1}{2!}\frac{d^2}{dx^2}\frac{(ax^2 + bc + c)}{x = \xi} = \frac{1}{2}.2a = a$$

UNIT 1 NUMERICAL DIFFERENTIATION

Structure **Page Nos.**

1.0 INTRODUCTION

Differentiation of a function is a fundamental and important concept in calculus. When the function, say f(x), is given explicitly, its derivatives $f'(x)$, $f''(x)$, $f'''(x)$,… etc. can be easily found using the methods of calculus. For example, if $f(x) = x^2$, we know that $f'(x) = 2x$, $f''(x) = 2$ and all the higher order derivatives are zero. However, if the function is not known explicitly but, instead, we are given a table of values of f(x) corresponding to a set of values of x, then we cannot find the derivatives by using methods of calculus. For instance, if $f(x_k)$ represents distance travelled by a car in time x_k, k = 0, 1, 2, …seconds, and we require the velocity and acceleration of the car at any time x_k, then the derivatives $f'(x)$ and $f''(x)$ representing velocity and acceleration respectively, cannot be found analytically. Hence, the need arises to develop methods of differentiation to obtain the derivative of a given function f(x), using the data given in the form of a table, where the data might have been formed as a result of scientific experiments.

Numerical methods have the advantage that they are easily adaptable on calculators and computers. These methods make use of the interpolating polynomials, which we discussed in earlier block. We shall now discuss, in this unit, a few numerical differentiation methods, namely, the method based on undetermined coefficients, methods based on finite difference operators and methods based on interpolation.

1.1 OBJECTIVES

After going through this unit you should be able to:

- explain the importance of the numerical methods over the methods of calculus;
- use the method of undetermined coefficients and methods based on finite difference operators to derive differentiation formulas and obtain the derivative of a function at step points, and
- use the methods derived from the interpolation formulas to obtain the derivative of a function at off step points.

1.2 METHODS BASED ON UNDETERMINED COEFFICIENTS

Earlier, we introduced to you the concepts of round-off and truncation errors. In the derivation of the methods of numerical differentiation, we shall be referring to these errors quite often. Let us first quickly recall these concepts before going further.

Round-off Error

The representation in a computer system, of a non-integer number is generally imprecise. This happens because space allotted to represent a number contains only fixed finite number of bits. But representation may either require more than finite number of allotted bits or even may require infinite number of bits. The situation is similar to the case when we attempt to represent the number 1/3 as a decimal number .333……

Any pre-assigned finite number of digits will fall short for representing 1/3 exactly as a sequence of decimal digits. Suppose, we are allotted 4 digits to represent (the fractional part) of a number, then the error in the representation is $.0000333\ldots = 10^{-4} \times .3333\ldots$

Thus $10^{-4} \times .3333$ is round-off error in the representation of the number 1/3 using 4 decimal digits on a paper. The numbers in the computer systems are represented in binary form, generally using floating point representation. The error in representation of a number due to fixed finite space allotted to represent it, is called round-off error.

Truncation Error

Truncation error in the values of a function arise due to the method used for computing the values of the function. Generally, the method giving rise to truncation error is a infinite process, i.e., involves steps.

However, as it is impossible to execute infinitely many steps, therefore, the process has to be truncated after finite number of steps. But the truncation of an infinite process of calculation, to finite number of steps, leads to error in the value of the function at a point. Such an error is called truncation error.

For example, a method of computing the value of the function

$$f(x) = e^x$$

at a point x is given by

$$f(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots\ldots$$

But, as it is impossible to execute the infinitely many steps of computing and adding $\frac{x^n}{n!}$ for all n, the process has to be truncated after finite number of steps say after computing and adding $\frac{x^3}{3!}$. Then the error in the value of f(x) is given by $\frac{x^4}{4!} + \frac{x^5}{5!} + \ldots\ldots$ and is called truncation error.

We may note the difference between round-off error and truncation error

The round-off error in the representation of a number which possibly requires infinite space for its exact representation. On the other hand truncation error in the computed value of a function at a point, is due to the fact that the computation process for computing the value may have infinite steps. But, as infinite number of steps can not be executed, hence we are compelled to stop the process after some finite number of *steps*. But this truncation of an infinite process after finite number of steps leads to an error called truncation error.

Definition: Let f(h) be the exact analytical value of a given function obtained by using an analytical formula and f_h be the approximate value obtained by using a

numerical method. If the error $f(h)-f_h = C\,h^{p+1}$, where C is a constant, then p is known as the **order of the numerical method,** and is denoted by $O(h^p)$.

The Method

Let us consider a function f (x), whose values are given at a set of tabular points. For developing numerical differentiation formulas for the derivatives f′ (x), f″ (x), …at a point $x = x_k$, we express the derivative f^q (x), $q \geq 1$, as a linear combination of the values of f(x) at an arbitrarily chosen set of tabular points. Here, we assume that the tabular points are equally spaced with the step-length h, i.e., various step (nodal) points are $x_m = x_0 \pm mh$, m = 0, 1, … etc. Then we write

$$h^q f^q(x_k) = \sum_{m=-s}^{n} \gamma_m f_{k+m}, \qquad (1)$$

where γ_i for i = −s, −s+1,…….,n are the unknowns to be determined and f_{k+m} denotes $f(x_k+mh)$. For example, when s = n = 1 and q = 1, Eqn. (1) reduces to

$$hf'(x_k) = \gamma_{-1}f_{k-1} + \gamma_0 f_k + \gamma_1 f_{k+1}.$$

Similarly, when s =1, n =2 and q = 2, we have

$$h^2 f''(x_k) = \gamma_{-1}f_{k-1} + \gamma_0 f_k + \gamma_1 f_{k+1} + \gamma_2 f_{k+2}.$$

Now suppose we wish to determine a numerical differentiation formula for $f^q(x_k)$ of order p using the method of undetermined coefficients. In other words, we want our formula to give the exact derivative values when f(x) is a polynomial of degree ≤ p, that is, for $f(x) = 1, x, x^2, x^3, \ldots, x^p$. We then get p+1 equations for the determination of the unknowns γ_i, i = −s, −s + 1, …, n. You know that if a method is of order p, then its truncation error (TE) is of the form $Ch^{p+1} f^{(p+1)}(\alpha)$, for some constant C. This implies that if $f(x) = x^m$, m = 0, 1, 2, …., p then the method gives exact results since

$$\frac{d^{p+1}}{dx^{p+1}}(x^m) = 0, \text{ for } m = 0, 1, .., p.$$

Let us now illustrate this idea to find the numerical differentiation formula of $O(h^4)$ for f″ (x_k).

Derivation of $O(h^4)$ formula for $f''(x_k)$

Without loss of generality, let us take $x_k = 0$. We shall take the points symmetrically, that is, $x_m = mh$; m = 0, ± 1, ± 2.

Let $f_{-2}, f_{-1}, f_0, f_1, f_2$ denote the values of f(x) at x = −2h, −h, 0, h, 2h respectively. In this case the formula given by Eqn. (1) can written as

$$h^2 f''(0) = \gamma_{-2}f_{-2} + \gamma_{-1}f_{-1} + \gamma_0 f_0 + \gamma_1 f_1 + \gamma_2 f_2 \qquad (2)$$

As there are five unknown to be determined let us make the formula exact for f(x) = 1, x, x^2, x^3, x^4. Then, we have

f(x) = 1, f″ (0) = 0; $f_{-2} = f_{-1} = f_0 = f_1 = f_2 = 1$

f(x) = x, f″ (0) = 0, $f_{-2} = -2h$; $f_{-1} = -h$; $f_0 = 0$; $f_1 = h$; $f_2 = 2h$;

$f(x) = x^2$, f″(0) = 2, $f_{-2} = 4h^2 = f_2$; $f_{-1} = h^2 = f_1$; $f_0 = 0$;

$f(x) = x^3$, $f''(0) = 0$, $f_{-2} = -8h^3$; $f_{-1} = -h^3$; $f_0 = 0$; $f_1 = h^3$, $f_2 = 8h^3$

$f(x) = x^4$, $f''(0) = 0$; $f_{-2} = 16h^4 = f_2$; $f_{-1} = h^4 = f_1$; $f_0 = 0$, (3)

where $f_i = f(ih)$, e.g., if $f(x) = x^3$ then $f_2 = f(2h) = (2h)^3 = 8h^3$.

Substituting these values in Eqn. (2), we obtain the following set of equations for determining γ_m, $m = 0, \pm 1, \pm 2$.

$$\begin{aligned}
&\gamma_{-2} + \gamma_{-1} + \gamma_0 + \gamma_1 + \gamma_2 = 0\\
&-2\gamma_{-2} - \gamma_{-1} + \gamma_1 + 2\gamma_2 = 2\\
&4\gamma_{-2} + \gamma_{-1} + \gamma_1 + 4\gamma_2 = 2\\
&-8\gamma_{-2} - \gamma_{-1} + \gamma_1 + 8\gamma_2 = 0\\
&16\gamma_{-2} + \gamma_{-1} + \gamma_1 + 16\gamma_2 = 0
\end{aligned} \qquad (4)$$

Thus we have a system of five equations for five unknowns. The solution of this system of Eqns. (4) is

$\gamma_{-2} = \gamma_2 = -1/12$; $\gamma_{-1} = \gamma_1 = 16/12$; $\gamma_0 = 30/12$;

Hence, the numerical differentiation formula of $O(h^4)$ for $f''(0)$ as given by Eqn. (2) is

$$f''(0) \approx f''_0 = \frac{1}{12h^2}\left[-f_{-2} + 16f_{-1} - 30f_0 + 16f_1 - f_2\right] \qquad (5)$$

Now, we know that TE of the formula (5) is given by the first non-zero term in the Taylor expression of

$$f''(x_0) - \frac{1}{12h^2}\left[-f(x_0 - 2h) + 16f(x_0 - h) - 30f(x_0) + 16f(x_0 + h) - f(x_0 + 2h)\right] \qquad (6)$$

The Taylor series expansions give

$$f(x_0 - 2h) = f(x_0) - 2hf'(x_0) + 2h^2 f''(x_0) - \frac{4h^3}{3}f'''(x_0) + \frac{2h^4}{3}f^{IV}(x_0) - \frac{4h^5}{15}f^{V}(x_0) + \frac{4h^6}{45}f^{VI}(x_0) - \ldots\ldots$$

$$f(x_0 - h) = f(x_0) - hf'(x_0) + \frac{h^2}{2}f''(x_0) - \frac{h^3}{6}f'''(x_0) + \frac{h^4}{24}f^{IV}(x_0) - \frac{h^5}{120}f^{V}(x_0) + \frac{h^6}{720}f^{VI}(x_0) + \ldots\ldots$$

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2}f''(x_0) + \frac{h^3}{6}f'''(x_0) + \frac{h^4}{24}f^{IV}(x_0) + \frac{h^5}{120}f^{V}(x_0) + \frac{h^6}{720}f^{IV}(x_0) + \ldots\ldots$$

$$f(x_0 + 2h) = f(x_0) + 2hf'(x_0) + \frac{2h^2}{2}f''(x_0) + \frac{4h^3}{3}f'''(x_0) + \frac{2h^4}{3}f^{IV}(x_0) + \frac{4h^5}{15}f^{V}(x_0)$$

$$+\frac{4h^6}{45}f^{VI}(x_0)+.....$$

Substituting these expansions in Eqn. (6) and simplifying, we get the first non-zero term or the TE of the formula (5) as

$$TE = f''(x_0)-\frac{1}{12h^2}\left[-f(x_0-2h)+16f(x_0-h)-30f(x_{0)}+16f(x_0+h)-f(x_0+2h)\right]$$

$$= -\frac{h^6}{90}f^{VI}(\alpha), 0<\alpha<1.$$

You may now try the following exercise.

Ex.1) A differentiation rule of the form
$f'_0 = \alpha_0 f_0 + \alpha_1 f_1 + \alpha_2 f_2$
is given. Find α_0, α_1 and α_2 so that the rule is exact for polynomials of degree 2 or less.

You must have observed that in the numerical differentiation formula discussed above, we have to solve a system of equations. If the number of nodal points involved is large or if we have to determine a method of high order, then we have to solve a large system of linear equations, which becomes tedious. To avoid this, we can use finite difference operators to obtain the differentiation formulas, which we shall illustrate in the next section.

1.3 METHODS BASED ON FINITE DIFFERENCE OPERATORS

Recall that in Unit 4 of Block 2, we introduced the finite difference operators E, ∇, Δ, μ and δ. There we also stated the relations among various operators, e.g.

$$E = \Delta + 1$$
$$= (1 - \nabla)^{-1}$$

$$\mu = \frac{1}{2}(E^{1/2} + E^{-1/2})$$

$$l_n\ E = \log_e E$$

In order to construct the numerical differentiation formulas using these operators, we shall first derive relations between the differential operator D where $Df(x) = f'(x)$ and the various difference operators.

By Taylor series, we have

$$f(x + h) = f(x) + hf'(x) + h^2 f''(x) + ...$$

$$= [1 + hD + h^2D^2 + ...]\ f(x)$$
$$= e^{hD}\ f(x) \qquad (7)$$

Using, $Ef(x) = f(x+h)$, we obtain from Eqn. (7), the identity

$$E = e^{hD} \tag{8}$$

Taking logrithm of both sides, we get

$$hD = \ln E = \ln (1+\Delta) = \Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4} + ... \tag{9}$$

$$hD = \ln E = -\ln (1-\nabla) = \nabla + \frac{\nabla^2}{2} + \frac{\nabla^3}{3} - \frac{\nabla^4}{4} + ... \tag{10}$$

We can relate D with δ as follows:

We know that $\delta = E^{1/2} - E^{-1/2}$. Using identity (8), we can write

$$\delta f(x) = \left[e^{hD/2} - e^{-hD/2}\right] f(x) \; +$$

Hence, $\delta = 2\sinh (hD/2)$

or $hD = 2\sinh^{-1} (\delta/2)$ (11)

Similarly $\mu = \cosh (hD/2)$ (12)

Then, we also have $\mu\,\delta = 2 \sinh (h\,D/2) \cosh (h\,D/2) = \sinh (hD)$

or $hD = \sinh^{-1} (\mu\,\delta)$ (13)

and $\mu^2 = \cosh^2 (hD/2) = 1 + \sinh^2 (hD/2) = 1 + \frac{\delta^2}{4}$ (14)

Using the Maclaurin's expansion of $\sinh^{-1}x$, in relation (11), we can express hD as an infinite series in $\delta/2$.

Thus, we have

$$hD = 2\sinh^{-1} (\delta/2)$$

$$= \delta - \frac{1^2.\delta^3}{2^2 3!} + \frac{1^2 3^2.\delta^5}{2^4.5!} + \frac{1^2.3^2.5^2.\delta^7}{2^6.7!} + ... \tag{15}$$

Notice that this formula involves off-step points when operated on f(x). The formula involving only the step points can be obtained by using the relation (13), i.e.,

$$hD = \sinh^{-1} (\mu\,\delta)$$

$$= \mu\delta - \frac{1^2\mu^3\delta^3}{3!} + \frac{1^2.3^2.\mu^5\delta^5}{5!} - \frac{1^2.3^2.5^2.\mu^7\delta^7}{7!} + ... \tag{16}$$

Using relation (14) in Eqn. (16), we obtain

$$hD = \mu\left[\delta - \frac{\delta^3}{6} + \frac{\delta^5}{30} - \frac{\delta^7}{140} + ...\right] \tag{17}$$

Thus, Eqns. (9), (10) and (17) give us the relation between hD and various difference operators. Let us see how we can use these relations to derive numerical differentiation formulas for f'_k, f''_k etc.

We first derive formulas for f'_k. From Eqn. (9), we get

$$hDf(x_k) = hf'_k = \left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4} + \ldots\right) f_k$$

Thus forward difference formulas of $O(h)$, $O(h^2)$, $O(h^3)$ and $0(h^4)$ can be obtained by retaining respectively 1,2 3, and 4 terms of the relation (9) as follows:

$0(h)$ method : $hf'_k = \Delta f_k = f_{k+1} - f_k$ (18)

$0(h^2)$ method : $hf'_k = \Delta f_k = \frac{1}{2}\left(-f_{k+2} + 4f_{k+1} - 3f_k\right)$ (19)

$0(h^3)$ method : $hf'_k = \left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3}\right) f_k = \frac{1}{6}\left(2f_{k+3} - 9f_{k+2} + 18f_{k+1} - 11f_k\right)$ (20)

$0(h^4)$ method :

$$hf'_k = \left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4}\right) f_k = \frac{1}{12}\left(-3f_{k+4} + 16f_{k+3} - 36f_{k+2} + 48f_{k+1} - 25f_k\right) \quad (21)$$

TE of the formula (18) is

$$TE = f'(x_k) - \frac{1}{h}\left[f(x_{k+1}) - f(x_k)\right] = -\frac{h}{2} f''(\xi) \quad (22)$$

and that of formula (19) is

$$TE = f'(x_k) - \frac{1}{2h}\left[-f(x_{k+2}) + 4f(x_{k+1}) - 3f(x_k)\right]^2 = \frac{h^2}{3} f'''(\xi) \quad (23)$$

Similarly the TE of formulas (20) and (21) can be calculated. Backward difference formulas of $0(h)$ $0(h^2)$, $0(h^3)$ and $0(h^4)$ for f_k can be obtained in the same way by using the equality (10) and retaining respectively 1,2,3 or 4 terms. We are leaving it as an exercise for you to derive these formulas.

Ex. 2) Derive backward difference formulas for f'_k of $0(h)$, $0(h^2)$, $0(h^3)$ and $0(h^4)$.

Central difference formulas for f'_k can be obtained by using the relation (17), i.e.,

$$hf'_k = \mu\left(\delta - \frac{\delta^3}{6} + \ldots\right) f_k$$

Note that relation (17) gives us methods of $0(h^2)$ and $0(h^4)$, on retaining respectively 1 and 2 terms, as follows:

$0(h^2)$ method : $hf'_k = \frac{1}{2}\left(f_{k+1} - f_{k-1}\right)$ (24)

$0(h^4)$ method : $hf'_k = \frac{1}{12}\left(-f_{k-2} + 8f_{k-1} + 8f_{k+1} + f_{k+2}\right)$ (25)

We now illustrate these methods through an example.

Example 1: Given the following table of values of $f(x) = e^x$, find $f'(0.2)$ using formulas (18), (19), (24) and (25).

x	:	0.0	0.1	0.2	0.3	0.4
f(x)	:	1.000000	1.105171	1.221403	1.349859	1.491825

Using $(18), f'(0.2) = \dfrac{f(0.3) - f(0.2)}{0.1}$

or $$f'(0.2) = \frac{1.349859 - 1.221403}{0.1}$$

$$= 1.28456$$

$$TE = -\frac{h}{2} f''(0.2) = -\frac{1}{2} e^{0.2} = -0.061070$$

Actual error = $1.221402758 - 1.28456 = -0.063157$

Using $(19), f'(0.2) = \dfrac{1}{0.2}\left[-f(0.4) + 4f(0.3) - 3f(0.2)\right] = 1.21701$

$$TE = \frac{h^2}{3} f'''(0.2) = \frac{0.01}{3} e^{0.2} = 0.004071;$$

Actual error = 0.004393

Using $(24), f'(0.2) = \dfrac{1}{0.2}\left[f(0.3) - f(0.1)\right] = 1.22344$

$$TE = -\frac{h^2}{6} f'''(0.2) = -\frac{0.01}{6} e^{0.2} = -0.0020357$$

Actual error $= -0.002037$

Using $(25), f'(0.2) = \dfrac{1}{12}\left[-f(0.0) + 8f(0.1) - 8f(0.3) + f(0.4)\right] = 1.221399167$

$$TE = \frac{h^4 f^{VI}(0.2)}{30} = \frac{0.0001}{30} e^{0.2} = 0.4071 \times 10^{-5};$$

$$\text{Actual error} = 0.3591 \times 10^{-5}$$

Numerical differentiation formulas for f''_k can be obtained by considering

$$h^2 D^2 = \Delta^2 - \Delta^3 + \frac{11}{12}\Delta^4 - \frac{2}{3}\Delta^5 + \ldots \quad (26)$$

$$= \nabla^2 + \nabla^3 + \frac{11}{12}\nabla^4 + \frac{2}{3}\nabla^5 + \ldots \quad (27)$$

$$= \delta^2 - \frac{\delta^4}{12} + \frac{\delta^6}{90} - \ldots \quad (28)$$

We can write the forward difference methods of 0(h), $0(h^2)$, $0(h^3)$ and $0(h^4)$ for f''_k using Eqn. (26) and retaining 1,2,3 and 4 terms respectively as follows:

$$0(h)\text{ method: } h^2 f''_k = f_{k+2} - 2f_{k+1} + f_k \quad (29)$$

$$0(h^2)\text{ method: } h^2 f''_k = -f_{k+3} + 4f_{k+2} - 5f_{k+1} + 2f_k \quad (30)$$

$$0(h^3)\text{ method: } h^2 f''_k = \frac{1}{12}\left(11f_{k+4} - 56f_{k+3} + 114f_{k+2} - 104f_{k+1} + 35f_k\right) \quad (31)$$

$$0(h^4)\text{ method: } h^2 f''_k = \frac{1}{12}\left(-8f_{k+5} + 51f_{k+4} - 136f_{k+3} + 194f_{k+2} - 144f_{k+1} + 43f_k\right) \quad (32)$$

Backward difference formulas can be written in the same way by using Eqn. (27). Central difference formulas of $0(h^2)$ and $0(h^4)$ for f''_k are obtained by using Eqn. (28) and retaining 1 or 2 terms respectively as follows:

$$0(h^2)\text{ method: } h^2 f''_k = \left(f_{k-1} - 2f_k + f_{k+1}\right) \quad (33)$$

$$0(h^4)\text{ method: } h^2 f''_k = \frac{1}{12}\left(-f_{k-2} + 16f_{k-1} - 30f_k + 16f_{k+1} - f_{k+2}\right) \quad (34)$$

Let us consider an example,

Example 2: For the table of values of $f(x) = e^x$, given in Example 1, find $f''(0.2)$ using the formulas (33) and (34).

Solution: Using (33), $f''(0.2) = \frac{1}{0.01}\left[f(0.1) - 2f(0.2) + f(0.3)\right] = 1.2224$

$$\text{TE} = \frac{-h^2 f^{IV}(0.2)}{12} = \frac{-(0.01)e^{0.2}}{12} = -0.0010178$$

Actual error = – 0.0009972

Using Eqn. (34),

$$f''(0.2) = \frac{\left[-f(0.0) + 16f(0.1) - 30f(0.2) + 16f(0.3) - f(0.4)\right]}{0.12} = 1.221375$$

$$\text{TE} = \frac{h^4 f^{VI}(0.2)}{90} = 0.13571 \times 10^{-5}$$

Actual error = 0.27758×10^{-4}

And now the following exercises for you.

Ex. 3) From the following table of values find $f'(6.0)$ using an 0(h) formula and $f''(6.3)$ using an $0(h^2)$ formula.

x	:	6.0	6.1	6.2	6.3	6.4
f(x)	:	0.1750	– 0.1998	– 0.2223	– 0.2422	–0.2596

Ex. 4) Calculate the first and second derivatives of $1_n x$ at x = 500 from the following table. Use $0(h^2)$ forward difference method. Compute TE and actual errors.

x	:	500	510	520	530
f(x)	:	6.2146	6.2344	6.2538	6.2729

In Secs. 5.2 and 5.3, we have derived numerical differentiation formulas to obtain the derivative values at **nodal** points or **step points**, when the function values are given in the form of a table. However, these methods cannot be used to find the derivative values at off-step points. In the next section we shall derive methods which can be used for finding the derivative values at the off-step points as well as at steps points.

1.4 METHODS BASED ON INTERPOLATION

In these methods, given the values of f(x) at a set of points $x_0, x_1, \ldots x_n$ the general approach for deriving numerical differentiation formulas is to obtain the unique interpolating polynomial $P_n(x)$ fitting the data. We then differentiate this polynomial q times ($q \le n$), to get $P_n^{q}(x)$. The value $P_n^{q}(x_k)$ then gives us the approximate value of $f^{(q)}(x_k)$ where x_k may be a step point or an off-step point. We would like to point out here that even when the original data are known to be accurate i.e. $P_n(x_k) = f(x_k)$, k = 0, 1, 2, …, n, yet the derivative values may differ considerably at these points. The approximations may further deteriorate while finding the values at off-step points or as the order of the derivative increases. However, these disadvantages are present in every numerical differentiation formula, as in general, one does not know whether the function representing a table of values has a derivative at every point or not.

We shall first derive differentiation formulas for the derivatives using non-uniform nodal points. That is, when the difference between any two consecutive points is not uniform.

Non-Uniform Nodal Points

Let the data (x_k, f_k), k = 0,1, …, n be given at n+1 points where the step length $x_i - x_{i-1}$ may not be uniform.

In Unit 4 you have seen that the Lagrange interpolating polynomial fitting the data (x_k, f_k), k = 0, 1, …, n is given by

$$P_n(x) = \sum_{k=0}^{n} L_k(x) f_k \qquad (35)$$

where $L_k(x)$ are the fundamental Lagrange polynomials given by

$$L_k(x) = \frac{\pi(x)}{(x - x_k)\pi'(x_k)} \qquad (36)$$

and $\pi(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$ (37)

$$\pi'(x_k) = (x_k - x_0)(x_k - x_1)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_n) \qquad (38)$$

The error of interpolation is given by

$$E_n(x) = f(x) - P_n(x) = \frac{\pi(x)}{(n+1)!} f^{(n+1)}(\alpha), \quad x_0 < \alpha < x_n$$

Differentiating $P_n(x)$ = w.r.t. x, we obtain

$$P'_n(x) = \sum_{k=0}^{n} L'_k(x) f_k \tag{39}$$

and the error is given by

$$E'_n(x) = \frac{1}{(n+1)!}\left\{\pi'(x) f^{(n+1)}(\alpha) + \pi(x)\left(f^{(n+1)}(\alpha)\right)'\right\} \tag{40}$$

Since in Eqn. (40), the function $\alpha(x)$ is not known in the second term on the hand side, we cannot evaluate $E'_n(x)$ directly. However, since at a nodal point $\pi(x_k) = 0$, we obtain

$$E'_n(x_k) = \frac{\pi'(x_k)}{(n+1)!} f^{(n+1)}(\alpha) \tag{41}$$

If we want to obtain the differentiation formulas for any higher order, say qth $(1 \le q \le n)$ order derivative, then we differentiate $P_n(x)$, q times and get

$$f^{(q)}(x) \approx P_n^{(q)}(x) = \sum_{k=0}^{n} L_k^{(q)}(x) f_k \tag{42}$$

Similarly, the error term is obtained by differentiating $E_n(x)$, q times.
Let us consider the following examples.

Example 3: Find f′(x) and the error of approximation using Lagrange Interpolation for the data (x_k, f_k), k = 0, 1.

Solution: We know that $P_1(x) = L_0(x) f_0 + L_1(x) f_1$

Where $L_0(x) = \frac{x - x_1}{x_0 - x_1}$ and $L_1(x) = \frac{x - x_0}{x_1 - x_0}$

Now,

$$P'_1(x) = L'_0(x) f_0 + L'_1(x) f_1$$

and $L'_0(x) = \frac{1}{x_0 - x_1}, L'_1(x) = \frac{1}{x_1 - x_0}$

Hence, $f'(x) = P'_1(x) = \frac{f_0}{x_0 - x_1} + \frac{f_1}{x_1 - x_0} = \frac{(f_1 - f_0)}{(x_1 - x_0)}$ (43)

$$E'_1(x_0) = \frac{(x_0 - x_1)}{2} f''(\alpha) \text{ and } E'_1(x_1) = \frac{(x_1 - x_0)}{2} f''(\alpha), \; x_0 < \alpha < x_1.$$

Example 4: Find f′(x) and f″(x) given f_0, f_1, f_2 at x_0, x_1, x_2 respectively, using Lagrange interpolation.

Solution: By Language's interpolation formula

$$f(x) \approx P_2(x) = L_0(x) f_0 + L_1(x) f_1 + L_2(x) f_2$$

where,

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}; \qquad L'_0(x) = \frac{2x - x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)}$$

$$L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}; \qquad L'_1(x) = \frac{2x - x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)}$$

$$L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}; \qquad L'_2(x) = \frac{2x - x_0 - x_1}{(x_2 - x_0)(x_2 - x_1)}$$

$$\text{Hence, } f'(x) = P'_2(x) = L'_0(x) f_0 + L'_1(x) f_1 + L'_2(x) f_2$$

$$\text{and } P''_2(x) = L''_0(x) f_0 + L''_1(x) f_1 + L''_2(x) f_2$$

$$= \frac{2f_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{2f_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{2f_2}{(x_2 - x_1)}$$

Example 5: Given the following values of f(x) = In x, find the approximate value of f′ (2.0) and f ′′(2.0). Also find the errors of approximations.

x	:	2.0	2.2	2.6
f(x)	:	0.69315	0.78846	0.95551

Solution: Using the Lagrange's interpolation formula, we have

$$f'(x_0) = P'_2(x_0) = \frac{2x_0 - x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)} f_0 + \frac{x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)} f_1 + \frac{x_0 - x_1}{(x_2 - x_0)(x_2 - x_1)} f_2$$

$\therefore$ we get

$$f'(2.0) = \frac{4 - 2.2 - 2.6}{(2 - 2.2)(2 - 2.6)}(0.69315) + \frac{2 - 2.6}{(2.2 - 2)(2.2 - 2.6)}(0.78846)$$

$$+ \frac{2 - 2.2}{(2.6 - 2)(2.6 - 2.2)}(0.95551) = 0.49619$$

The exact value of f′ (2.0) = 0.5

Error is given by

$$E'_2(x_0) = \frac{1}{6}(x_0 - x_1)(x_0 - x_2) f'''(2.0)$$

$$= \frac{1}{6}(2.0 - 2.2)(2.0 - 2.6)(-0.25) = -0.005$$

Similarly,

$$f''(x_0) = 2\left[\frac{f_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{f_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{f_2}{(x_2 - x_0)(x_2 - x_1)}\right]$$

$$\therefore f''(2.0) = 2\left[\frac{0.69315}{(2-2.2)(2-2.6)} + \frac{0.78846}{(2.2-2)(2.2-2.6)} + \frac{0.95551}{(2.6-2)(2.6-2.2)}\right]$$

$$= -\ 0.19642$$

The exact value of $f''(2.0) = -\ 0.25$.

Error is given by

$$E''_2(x_0) = \frac{1}{3}(2x_0 - x_1 - x_2) f'''(2.0) + \frac{1}{6}(x_0 - x_1)(x_0 - x_2)\left[f^{IV}(2.0) + f^{IV}(2.0)\right]$$

$$= -\ 0.05166$$

You may now try the following exercise.

Ex.5) Use Lagrange's interpolation to find $f'(x)$ and $f''(x)$ at each of the values $x = 2.5; 5.0$ from the following table

x	:	1	2	3	4
f(x)	:	1	16	81	256

Next, let us consider the case of uniform nodal points.

Uniform Nodal Points

When the difference between any two consecutive points is the same, i.e., when we are given values of f(x) at equally spaced points, we can use Newton's forward or backward interpolation formulas to find the unique interpolating polynomial $P_n(x)$. We can then differentiate this polynomial to find the derivative values either at the nodal points or at off-step points.

Let the data (x_k, f_k), $k = 0, 1, \ldots, n$ be given at $(n+1)$ points where the step points x_k, $k = 0, 1, \ldots, n$ are equispaced with step length h. That is, we have $x_k = x_0 + kh$, $k = 1, 2, \ldots, n$.

You know that by Newton's forward interpolation formula

$$f(x) = P_n(x) = f_0 + \frac{(x - x_0)}{h}\Delta f_0 + \frac{(x - x_0)(x - x_1)}{2!h^2}\Delta^2 f_0 + \ldots$$

$$+ \frac{(x - x_0)(x - x_1)\ldots(x - x_{n-1})\Delta^n f_0}{n!h^n} \qquad (44)$$

with error

$$E_n(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_n)}{(n+1)!h^{n+1}}\Delta^{n+1} f(\alpha), \quad x_0 < \alpha < x_n. \qquad (45)$$

If we put $\frac{x - x_0}{h} = s$ or $x = x_0 + sh$, then Eqns. (44) and (45) reduce respectively to

$$f(s) = P_n(s) = f_0 + \frac{s\Delta f_0}{1!} + \frac{s(s-1)\Delta^2 f_0}{2!} + \frac{s(s-1)(s-2)\Delta^3 f_0}{3!} + \ldots$$

$$+ \frac{s(s-1)\ldots(s-n+1)\Delta^n f_0}{n!} \qquad (46)$$

and

$$E_n(x) = \frac{s(s-1)\ldots(s-n)}{(n+1)!} h^{(n+1)} f^{(n+1)}(\alpha) \qquad (47)$$

$$\frac{dP}{dx} = \frac{dp}{ds} \times \frac{ds}{dx} = \frac{1}{h} \cdot \frac{dP(s)}{ds}$$

$$\because\ dx = 0 + h\,ds \text{ or } \frac{ds}{dx} = \frac{1}{h}.$$

Now from (46) we get,

Differentiation of $P_n(x)$ w.r.t. x gives us

$$P'_n(x) = \frac{1}{h}\left[\Delta f_0 + \frac{(2s-1)}{2}\Delta^2 f_0 + \frac{(3s^2 - 6s + 2)}{6}\Delta^3 f_0 + \ldots\right] \qquad (48)$$

At $x = x_0$, we have $s = 0$ and hence

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{\Delta^2 f_0}{2} + \frac{\Delta^3 f_0}{3} + \frac{\Delta^4 f_0}{4} + \ldots\right]$$

which is same as formula (9) obtained in Sec. 5.3 by difference operator method. We can obtain the derivative at any step or off-step point by finding the value of s and substituting the same in Eqn. (48). The formula corresponding to Eqn. (48) in backward differences is

$$P'_n(x) = \frac{1}{h}\left[\nabla f_n + \frac{(2s+1)}{2}\nabla^2 f_n + \frac{(2s^2 + 6s + 2)}{6}\nabla^3 f_n + \ldots\right] \qquad (49)$$

where $x = x_n + sh$.

Formulas for higher order derivatives can be obtained by differentiating $P'_n(x)$ further and the corresponding error can be obtained by differentiating $E'_n(x)$.

Let us illustrate the method through the following examples:

Example 6: Find the first and second derivatives of f(x) at x = 1.1 from the following tabulated values.

x	:	1.0	1.2	1.4	1.6	1.8	2.0
f(x)	:	0.0000	0.1280	0. 5440	1.2960	2.4320	4.0000

Table 1

x	f(x)	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$	$\Delta^5 f(x)$
1.0	0.0					
		0.1280				
1.2	0.1280		0.2880			
		0.4160		0.0480		
1.4	0.5440		0.3360		0.0000	
		0.7520		0.0480		0.0000
1.6	1.2960		0.3840		0.0000	
		1.1360		0.0480		
1.8	2.4320		0.4320			
		1.5680				
2.0	4.0000					

Since, $x = x_0 + s\,h$, $x_0 = 1$, $h = 0.2$ and $x = 1.1$, we have $s = \dfrac{1.1-1}{0.2} = 0.5$

Substituting the value of s in formula (48), we get

$$f'(1.1) = \frac{1}{h}\left[\Delta f_0 - \frac{0.25}{6}\Delta^3 f_0\right] \qquad (50)$$

Substituting the values of Δf_0 and $\Delta^3 f_0$ in Eqn. (50) from Table 1, we get

$$f'(1.1) = 0.63$$

To obtain the second derivative, we differentiate formula (48) and obtain

$$f''(x) = P''(x) = \frac{1}{h}\left[\Delta^2 f_0 + (s-1)\Delta^3 f_0\right]$$

Thus $f''(1.1) = 6.6$

Note: If you construct a forward difference interpolating polynomial P(x), fitting the data given in Table 1, you will find that $f(x) = P(x) = x^3 - 3x + 2$. Also, $f'(1.1) = 6.3$, $f''(1.1) = 6.6$. The values obtained from this equation or directly as done above have to be same as the interpolating polynomial is unique.

Example 7: Find $f'(x)$ at $x = 0.4$ from the following table of values.

x	:	0.1	0.2	0.3	0.4	0.5
f(x)	:	1. 10517	1.22140	1.34986	1.49182	256

Solution: Since we are required to find the derivative at the right-hand end point, we will use the backward difference formula. The backward difference table for the given data is given by

Table 2

X	f(x)	$\nabla f(x)$	$\nabla^2 f(x)$	$\nabla^3 f(x)$
0.1	1.10517			
		0.11623		
0.2	1.22140		0.01223	
		0.12846		0.00127
0.3	1.34986		0.01350	
		0.14196		
0.4	1.49182			

Since $x_n = 0.4$, $h = 0.1$, $x = 0.4$, we get $s = 0$

Substituting the value of s in formula (49), we get

$$f'(0.4) = \frac{1}{4}\left[\Delta f_3 + \frac{1}{2}\Delta^2 f_3 + \Delta^3 f_3\right]$$

$$= \frac{1}{0.1}\left[0.14196 + \frac{0.0135}{2} + \frac{0.00127}{3}\right]$$

$$= 1.14913$$

How about trying a few exercises now ?

Ex.6) The position f(x) of a particle moving in a line at different point of time x_k is given by the following table. Estimate the velocity and acceleration of the particle at points x = 15 and 3.5

x	:	0	1	2	3	4
f(x)	:	–25	–9	0	7	15

Ex.7) Construct a difference table for the following data

x	:	1.3	1.5	1.7	1.9	2.1	2.3	2.5
f(x)	:	3.669	4.482	5.474	6.686	8.166	9.974	12.182

Taking h = 0.2, compute f′ (1.5) and the error, if we are given $f(x) = e^x$

Ex.8) Compute f″ (0.6) from the following table using $0(h^2)$ central difference formula with step lengths h = 0.4, 0.2, 0.1.

x	:	0.2	0.4	0.5	0.6	0.7	0.8	1.0
f(x)	:	1.420072	1.881243	2.128147	2.386761	2.657971	2.942897	3.559753

Ex. 9) Using central difference formula of $0(h^2)$ find f″ (0.3) from the given table

x	:	01	0.2	0.3	0.4	0.5
f(x)	:	0.091	0.155	0.182	0.171	0.130

1.5 SUMMARY

In this unit we have covered the following:

1) If a function f(x) is not known explicitly but is defined by a table of values of f(x) corresponding to a set of values of x, then its derivatives can be obtained by numerical differentiation methods.

2) Numerical differentiation formulas using

   (i) the method of undetermined coefficients and
   (ii) methods based on finite difference operators can be obtained for the derivatives of a function at nodal or step points when the function is given in the form of table.

3) When it is required to find the derivative of a function at off-step points then the methods mentioned in (2) above cannot be used. In such cases, the methods derived from the interpolation formulas are useful.

1.6 SOLUTIONS/ANSWERS

E1) Let $f'(x) = \alpha_0 f_0 + \alpha_1 f_1 + \alpha_2 f_2$. Setting $f(x) = 1, x, x^2$ we obtain

$$\alpha_0 + \alpha_1 + \alpha_2 = 0$$

For f(x) = x, $f_0 = 0$, $f_1 = h$ and $f_2 = 2h$, therefore,

$$(\alpha_1 + 2\alpha_2)h = 1$$

For $f(x) = x^2$, $f_0 = 0$, $f_1 = h^2$ and $f_2 = 4h^2$ therefore,

$$(\alpha_1 + 4\alpha_2)h^2 = 0, \text{ hence, } \alpha_1 + 4\alpha_2 = 0$$

Solving we obtain $\alpha_0 = -\frac{3}{2h}$, $\alpha_1 = \frac{2}{h}$ and $\alpha_2 = -\frac{1}{2h}$.

Hence, $f'_0 = \left(\frac{-3f_0 + 4f_1 - f_2}{2h}\right)$

E2) 0(h) method: $hf'_k = (f_k - f_{k-1})$

0(h^2) method: $hf'_k = \left(\frac{3f_k - 2f_{k-1} + f_{k-2}}{2}\right)$

0(h^3) method: $hf'_k = \left(\frac{11f_k - 22f_{k-1} + 9f_{k-2} - 6f_{k-3}}{6}\right)$

0(h^4) method: $hf'_k = \left(\frac{25f_k - 56f_{k-1} + 36f_{k-2} - 24f_{k-3} + 3f_{k-4}}{12}\right)$

E3) Using formula (18), we have

$$f'(6.0) = \left[\frac{f(6.1) - f(6.0)}{0.1}\right] = -3.7480$$

Using formula (33).

$$f''(6.3)=\left[\frac{f(6.4)-2f(6.3)+f(6.2)}{(0.1)^2}\right]=0.25$$

E4) Using formula (19), we have

$$f'(500)=\left[\frac{-3f(500)+4f(510)-f(520)}{2h}\right]=0.002$$

Using (30), we have

$$f''(500)=\left[\frac{2f(500)-5f(510)+4f(520)-f(530)}{h^2}\right]=-0.5\times10^{-5}$$

Exact value $f'(x) = 1/x = 0.002$; $f'''(x) = -1/x^2 = -0.4\times 10^{-5}$

Actual error in $f'(500)$ is 0, whereas in $f''(500)$ it is 0.1×10^{-5}. Truncation error in $f'(x)$ is $\frac{-h^2f'''}{3}=-5.33\times10^{-7}$ and in $f''(x)$ it is $\frac{11h^2f^{IV}}{12}=8.8\times10^{-9}$

E5) In the given problem $x_0 = 1$, $x_1 = 2$, $x_2 = 3$, $x_3 = 4$ and $f_0 = 1$, $f_1 = 16$, $f_2 = 81$ and $f_3 = 256$.

Constructing the Lagrange fundamental polynomials, we get

$$L_0(x)=-\left(\frac{x^3-9x^2+26x-24}{6}\right);L_1(x)=\left(\frac{x^3-8x^2+19x-12}{2}\right)$$

$$L_2(x)=-\left(\frac{x^3-7x^2+14x-8}{2}\right);L_3(x)=\left(\frac{x^3-6x^2+11x-6}{6}\right)$$

$$P_3(x)=L'_0(x)f_0+L'_1(x)f_1+L'_2(x)f_2+L'_3(x)f_3$$

$$P'_3(x)=L'_0(x)f_0+L'_1(x)f_1+L'_2(x)f_2+L'_3(x)f_3$$

$$P''_3(x)=L''_0(x)f_0+L''_1(x)f_1+L''_2(x)f_2+L''_3(x)f_3$$

We obtain after substitution,

$$P'_3(2.5)=62.4167;P''_3(2.5)=79;P'_3(5)=453.667;P''_3(5)=234.$$

The exact values of $f'(x)$ and $f''(x)$ are (from $f(x) = x^4$)

$f'(2.5) = 62.5$, $f'(5) = 500$; $f''(2.5) = 75$; $f''(5) = 300$.

E6) We are required to find $f'(x)$ and $f''(x)$ at $x = 1.5$ and 3.5 which are off-step points. Using the Newton's forward difference formula with $x_0 = 0$, $x = 1.5$, $s = 1.5$, we get $f'(1.5) = 8.7915$ and $f''(1.5) = -4.0834$.

Using the backward difference formula with $x_n = 4$, $x = 3.5$, $s = -0.5$, we get $f'(3.5) = 7.393$ and $f''(3.5) = 1.917$.

E7) The difference table for given problem is:

X	f(x)	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
1.3	3.669				
		0.813			
1.5	4.482		0.179		
		0.992		0.41	
1.7	5.474		0.220		0.007
		1.212		0.48	
1.9	6.686		0.268		0.012
		1.480		0.060	
2.1	8.166		0.328		0.012
		1.8.08		0.072	
2.3	9.974		0.400		
		2.208			
2.5	12.182				

Taking $x_0 = 1.5$ we see that s = 0 and we obtain from the interpolation formula

$$f'(1.5) = \frac{1}{h}\left[\Delta f_0 - \frac{\Delta^2 f_0}{2} + \frac{\Delta^3 f_0}{3} - \frac{\Delta^4 f_0}{4} + \ldots\right]$$

$$= \left[0.992 - \frac{0.220}{2} + \frac{0.048}{3} - \frac{0.012}{4}\right] = 4.475$$

Exact value is $e^{1.5} = 4.4817$ and error is = 0.0067

E8) Use the $0(h^2)$ formula (33). With h = 0.1 f ″ (0.6) = 1.2596, h = 0.2, f ″ (0.6) = 1.26545, h = 0.4, f ″ (0.6) = 1.289394.

E9) Using (24) with h = 0.1, we have f ′ (0.3) = –3.8

and with h = 0.2, f ″ (0.3) = –3.575

UNIT 2 NUMERICAL INTEGRATION

Structure **Page Nos**

2.0 INTRODUCTION

In Unit 1, we developed methods of differentiation to obtain the derivative of a function f(x), when its values are not known explicitly, but are given in the form of a table. In this unit, we shall derive numerical methods for evaluating the definite integrals of such functions f(x). You may recall that in calculus, the definite integral of f(x) over the interval [a, b] is defined as

$$\int_a^b f(x)\,dx = \lim_{h \to 0} R[h]$$

where R[h] is the left-end Riemann sum for n subintervals of length $h = \frac{(b-a)}{n}$ and is given by

$$R[h] = \sum_{k=0}^{n-1} h\, f(x_k)$$

for the nodal points $x_0, x_1, \ldots, x_n$, where $x_k = x_0 + kh$ and $x_0 = a$, $x_n = b$.

The need for deriving accurate numerical methods for evaluating the definite integral arises mainly, when the integrand is either

i) a complicated function such as $f(x) = e^{-x^2}$ or $f(x) = \frac{\sin(x)}{x}$ etc. which have no anti-derivatives expressible in terms of elementary functions, or

ii) when the integrand is given in the form of tables.

Many scientific experiments lead to a table of values and we may not only require an approximation to the function f(x) but also may require approximate representations of the integral of the function. Also, for functions the integrals of which can be calculated analytically, analytical evaluation of the integral may lead to transcendental, logarithmic or circular functions. The evaluation of these functions for a given value of x may not be accurate. This motivates us to study numerical integration methods which can be easily implemented on calculators.

In this unit we shall develop numerical integration methods where in the integral is approximated by a linear combination of the values of the integrand i.e.

$$\int_a^b f(x)\,dx = \beta_0 f(x_0) + \beta_1 f(x_1) + \ldots\ldots\ldots + \beta_n f(x_n) \qquad (1)$$

where $x_0, x_1, \ldots\ldots x_n$ are the points which divide the interval [a, b] into n sub-intervals and $\beta_0, \beta_1, \ldots\ldots \beta_n$ are the weights to be determined. We shall discuss in this unit, a few techniques to determine the unknowns in Eqn. (1).

2.1 OBJECTIVES

After going through this unit you should be able to:

- state the basic idea involved in numerical integration methods for evaluating the definite integral of a given function;
- use numerical integration method to find the definite integral of a function f(x) whose values are given in the form of a table;
- use trapezoidal and Simpson's rules of integration to integrate such functions and find the errors in these formulas.

2.2 METHODS BASED ON INTERPOLATION

In Block 2, you have studied interpolation formulas, which fits the given data (x_k, f_k), k = 0, 1, 2, ……., n. We shall now see how these interpolation formulas can be used to develop numerical integration methods for evaluating the definite integral of a function which is given in a tabular form. The problem of numerical integration is to approximate the definite integral as a linear combination of the values of f(x) in the form

$$\int_a^b f(x)\,dx \approx \sum_{k=0}^{n} \beta_k f_k \qquad (2)$$

where the n + 1 distinct points x_k, k = 0, 1, 2, ……., n are called the nodes or abscissas which divide the interval [a, b] into n sub-intervals with $(x_0 < x_1, < x_2, \ldots\ldots < x_{n-1} < x_n)$ and β_k, k =0, 1, ….n, are called the weights of the integration rule or quadrature formula. We shall denote the exact value of the definite integral by I and denote the rule of integration by

$$I_h[f] = \sum_{k=0}^{n} \beta_k f_k \qquad (3)$$

The error of approximating the integral I by $I_h[f]$ is given by

$$E_h[f] = \int_a^b f(x)dx - \sum_{k=0}^{n} \beta_k f_k \qquad (4)$$

The order of the integration method (3) is defined as follows:

Definition : An integration method of the form (3) is said to be of order p if it produces exact results for all polynomials of degree less than or equal to p.

In Eqn. (3) we have 2n+2 unknowns viz., n + 1 nodes x_k and the n + 1 weights, β_k and the method can be made exact for polynomials of degree $\leq 2n + 1$. Thus, the method of the form (3) can be of maximum order 2n + 1. But, if some of the values are prescribed in advance, then the order will be reduced. If all the n + 1 nodes are prescribed, then we have to determine only n + 1 weights and the corresponding method will be of maximum order n.

We first derive the numerical method based on Lagrange's interpolation.

2.2.1 Methods Using Lagrange's Interpolation

Suppose we are given the n + 1 abscissas x_k's and the corresponding values f_k's are known that the unique Lagrange's interpolating polynomial $P_{nn}(x)$ of degree ≤ n, satisfying the interpolatory conditions $P_n(x_k) = f(x_k)$, k = 0, 1, 2, ….. n, is given by

$$f(x) \approx P_n(x) = \sum_{k=0}^{n} L_k(x) f_k \tag{5}$$

with the error of interpolation

$$E_{n+1}[P_n(x)] = \frac{\pi(x)}{(n+1)!} f^{n+1}(\alpha) \text{ , with } x_0 < \alpha < x_n \tag{6}$$

where $L_k(x) = \dfrac{(x-x_0)(x-x_1)....(x-x_{k-1})(x-x_{k+1})....(x-x_n)}{(x_k-x_0)(x_k-x_1).....(x_k-x_{k-1})(x_k-x_{k+1})....(x_k-x_n)}$

and $\pi(x) = (x-x_0)(x-x_1)...(x-x_n)$.

We replace the function f(x) in the definite integral (2) by the Lagrange interpolating polynomial $P_n(x)$ given by Eqn. (5) and obtain

$$I_n[f] = \int_a^b P_n(x)dx = \sum_{k=0}^{n} \int_a^b L_k(x) f_k dx = \sum_{k-0}^{n} \beta_k f_k \tag{7}$$

where $\beta_k = \int_a^b L_k(x)\,dx.$ (8)

The error in the integration rule is

$$E_n[f] = \int_a^b E_{n+1}[P_n(x)]\,dx = \int_a^b \frac{\pi(x)}{(n+1)!} f^{n+1}(\alpha)dx \tag{9}$$

We have

$$|E_n[f]| \le \frac{M_{n+1}}{(n+1)!} \int_a^b |\pi(x)|dx \tag{10}$$

where $M_{n+1} = \max_{x_0 < x < x_n} \left| f^{n+1}(x) \right|$

Let us consider now the case when the nodes x_k 's are equispaced with $x_0 = a$, $x_n = b$ with the length of each subinterval as $h = \dfrac{b-a}{n}$. The numerical integration methods given by (7) are then known as **Newton-Cotes formulas** and the weights β_k's given by (8) are known as **Cotes numbers.** Any point x ε [a, b] can be written as $x = x_0 + sh$.

With this substitution, we have

$\pi(x) = h^{n+1} s\,(s-1)\,(s-2)\,(s-3)\,\ldots\ldots(s-n)$

$$L_k(x) = \frac{(-1)^{n-k} s(s-1)....(s-k+1)(s-k-1).....(s-n)}{k!(n-k)!} \qquad (11)$$

Using $x = x_0 + sh$ and changing the variable of integration from x to s, we obtain

$$\beta_k = \frac{(-1)^{n-k}}{k!(n-k)!} h \int_0^n s(s-1)(s-2).....(s-k+1)(s-k-1).....(s-n)\, ds \qquad (12)$$

and $$\left| E_n[f] \right| \le \frac{h^{n+2} M_{n+1}}{(n+1)!} \int_0^n s(s-1)(s-2).....(s-n) ds \qquad (13)$$

We now derive some of the Newton Cotes formulas viz. trapezoidal rule and Simpson's rule by using first and second degree Lagrange polynomials with equally spaced nodes.

Trapezoidal Rule

When $n = 1$, we have $x_0 = a$, $x_1 = b$ and $h = b - a$. Using Eqn. (12) the Cotes numbers can be found as

$$\beta_0 = -h \int_0^1 (s-1) ds = \frac{h}{2};$$

and $$\beta_1 = h \int_0^1 s\, ds = \frac{h}{2}.$$

Substituting the values of β_0 and β_1 in Eqn. (7), we get

$$I_T[f] = \frac{h}{2}[f_0 + f_1] \qquad (14)$$

The error of integration is

$$\left| E_T[f] \right| \le \frac{h^3}{2} M_2 \int_0^1 \left| s(s-1) \right| ds = \left| -\frac{h^3}{12} M_2 \right| = \frac{h^3}{12} M_2$$

Where $M_2 = \max_{x_0 < x < x_1} \left| f''(x) \right|$

Thus, by trapezoidal rule, $\int_a^b f(x)\, dx$ is given by

$$\boxed{I[f] = \frac{h}{2}(f_0 + f_1) - \frac{h^3}{12} M_2}$$

The reason for calling this formula the trapezoidal rule is that geometrically when f(x) is a function with positive values then $h(f_0 + f_1)/2$ is the area of the trapezium when height $h = b - a$ and parallel sides as f_o and f_1. This is an approximation to the area under the curve $y = f(x)$ above the x-axis bounded by the ordinates $x = x_1$ (see Fig. 1.)

Since the error given by Eqn. (15) contains the second derivative, trapezoidal rule integrates exactly polynomials of degree ≤ 1.

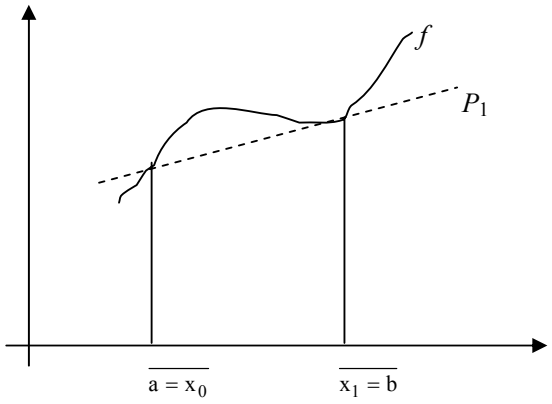


Fig. 1

Let us now consider an example.

Example 1: Find the approximate value of

$$I = \int_0^1 \frac{dx}{1+x}$$

using trapezoidal rule and obtain a bound for the error. The exact value of $I = \ln_2 = 0.693147$ correct to six decimal places.

Solution: Here $x_0 = 0$, $x_1=1$, and $h = 1 - 0 = 1$. Using Eqn. (14), we get

$$I_T\ [f] = \frac{1}{2}\left(1+\frac{1}{2}\right) = 0.75$$

Actual error = $0.75 - 0.693147 = 0.056853$.

The error in the trapezoidal rule is given by

$$|E_T|[f] \leq \frac{1}{12}\max\left|\frac{2}{(1+x)^3}\right| = \frac{1}{6} = 0.166667$$

Thus, the error bound retain is much greater than the actual error.

We now derive the Simpson's rule.

Simpson's Rule

For $n = 2$, we have $h = \frac{b-a}{2}$, $x_0 = a$, $x_1 = \frac{a+b}{2}$ and $x_2 = b$.

From (12), we find the Cotes numbers as

$$\beta_0 = \frac{h}{2}\int_0^2 (s-1)(s-2)\,ds = \frac{h}{3}$$

$$\beta_1 = h\int_0^2 s(s-2)\,ds = \frac{4h}{3}, \beta_2 = \frac{h}{2}\int_0^2 s(s-1)\,ds = \frac{h}{3}.$$

Eqn. (7) in this case reduces to

$$I_S[f] = \frac{h}{3}[f_0 + 4f_1 + f_2] \qquad (17)$$

Eqn. (17) is the Simpson's rule for approximating $I = \int_a^b f(x)\,dx$

The magnitude of the error of integration is

$$|E_2[f]| \le \frac{h^4 M_3}{3!}\int_0^2 |s(s-1)(s-2)|\,ds$$

$$= \frac{h^4 M_3}{3!}\left[\int_0^1 s(s-1)(s-2)\,ds + \int_1^2 s(s-1)(s-2)\,ds\right]$$

$$= \frac{h^4 M_3}{3!}\left[\left(\frac{s^4}{4} - s^3 + s^2\right)_0^1 + \left(\frac{s^4}{4} - s^3 + s^2\right)_1^2\right]$$

$$= \frac{h^4 M_3}{3!}\left[\frac{1}{4} - \frac{1}{4}\right] = 0$$

This indicates that Simpson's rule integrates polynomials of degree 3 **also** exactly. Hence, we have to write the error expression (13) with n = 3. We find

$$|E_S[f]| \le \frac{h^5 M_4}{24}\int_0^2 |s(s-1)(s-2)(s-3)|\,ds$$

$$= \frac{h^5 M_4}{24}\left[\int_0^1 s(s-1)(s-2)(s-3)\,ds + \int_1^2 s(s-1)(s-2)(s-3)\,ds\right] \qquad (18)$$

$$= \frac{-h^5 M_4}{90}$$

where $M_4 = \max_{x_0 < x < x_2} |f^{IV}(X)|$

Since the error in Simpson's rule contains the fourth derivative, Simpson's rule integrates exactly all polynomials of degree ≤ 3.

Thus, by Simpson's rule, $\int_a^b f(x)\,dx$ is given by

$$\boxed{I_S[f] = \frac{h}{3}[f_0 + 4f_1 + f_2] - \frac{h^5}{90}M_4}$$

Geometrically, $\frac{h}{3}[f_0 + 4f_1 + f_2]$ represents the area bounded by the quadratic curve passing through (x_0 , f_0), (x_1 , f_1) and (x_2 , f_2) above the x-axis and lying between the ordinates $x = x_0$, $x = x_2$ (see figure. 2).

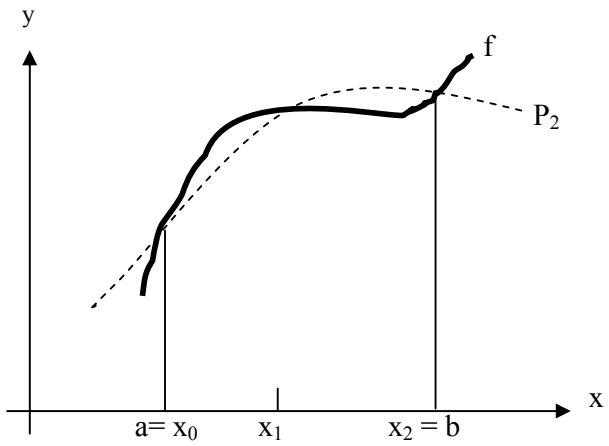


Fig. 2

In case we are given only one tabulated value in the interval [a, b], then h = b – a, and the interpolating polynomial of degree zero is $P_0(x) = f_k$. In this case, we obtain the rectangular integration rule given by

$$\boxed{I_R[f] = \int_a^b f_k \, dx \approx h f_k} \qquad (19)$$

The error in the integration rule is obtained from Eqn. (13) as

$$E_R|f| \le \frac{h^2 M_1}{2} \qquad (20)$$

where $M_1 = \max_{a < x < b} |f'(x)|$

If the given tabulated value in the interval [a, b] is the value at the mid-point then we have $x_k = \frac{(a+b)}{2}$, and $f_k = f_{k+\frac{1}{2}}$. In this case h = b – a and we obtain the integration rule as

$$\boxed{M\, f = \int_a^b \; {}_{k+\frac{1}{2}} \, dx \qquad {}_{k+\frac{1}{2}}} \qquad (21)$$

Rule (21) is called the mid-point rule. The error in the rule calculated from (13) is

$$E_M[f] = \frac{h^2}{2}\int_{-1/2}^{1/2} s\,ds = 0$$

This shows that mid-point rule integrates polynomials of degree one exactly. Hence the error for the mid-point rule is given by

$$E_M[f] \le \frac{h^3 M_2}{2}\int_{-1/2}^{1/2} s(s-1)\,ds = \frac{h^3 M_2}{24} \qquad (22)$$

where $M_2 = \max_{a < x < b} |f''(x)|$ and $h = b - a$

We now illustrate these methods through an example.

Example 2 : Evaluate $\int_0^1 e^{-x^2}\,dx$, using

a) Rectangular rule b) mid-point rule c) trapezoidal rule and d) Simpson's rule. If the exact value of the integral is 0.74682 correct to 5 decimal places, find the error in these rules.

Solution: The values of the function $f(x) = e^{-x^2}$ at $x = 0,\ 0.5$ and 1 are

f(0) = 1, f(0.5) = 0.7788, f(1) = 0.36788

Taking h = 1 and using

a) $I_R[f] = hf_{0,}$ we get $I_R[f] = 1.$

b) $I_M[f] = hf_{1/2,}$ we get $I_M[f] = 0.7788$

c) $I_T[f] = \frac{h}{2}[f_0 + f_1]$ we get $I_T[f] = \frac{1}{2}(1 + 0.36788) = 0.68394.$

Taking h = 0.5 and using Simpson's rule, we get

d) $I_S[f] = \frac{h}{3}[f_0 + 4f_1 + f_3]$

$= \frac{h}{3}[f(0) + 4f(0.5) + f(1)]$

= 0.74718

Exact value of the integral is 0.74682.

The errors in these rules are given by

$E_R[f] = -0.25318, E_M[f] = -0.03198$

$E_T[f] = 0.06288, E_s[t] = -0.00036.$

You may now try the following exercise:

Ex.1) Use the trapezoidal and Simpson's rule to approximate the following integrals. Compare the approximations to the actual value and find a bound for the error in each case.

a) $\int_1^2 \ln x\,dx$

b) $\int_0^{0.1} x^{1/3}\,dx$

c) $\int_0^{\pi/4} \tan x\,dx$

We now derive integration methods using Newton's forward interpolation formula.

2.2.2 Methods Using Newton's Forward Interpolation

Let the data be given at equi-spaced nodal points $x_k = x_0 + sh, s = 0,1,2,.....,n,$
Where $x_0 = a$ and $x_n = x_0 + nh = b$.

The step length is given by $h = \dfrac{b-a}{n}$.

The Newton's forward finite difference interpolation formula interpolating this data is given by

$$f(x) \approx P_n(x) = f_0 + s\Delta f_0 + s(s-1)\frac{\Delta^2 f_0}{2} + ... + \frac{s\,(s-1)(s-2)...(s-n+1)\Delta^n f_0}{n!} \qquad (23)$$

with the error of interpolation

$$E_n[f] = \frac{h^{n+1} s(s-1)(s-2)...(s-n)}{(n+1)!} f^{n+1}(\alpha)$$

Integrating both sides of Eqn. (23) w.r.t. x between the limits a and b, we can approximate the definite integral I by the numerical integration rule

$$I_h[f] = \int_a^b P_n(x)\,dx = h\int_0^1 \left[f_0 + s\,\Delta f_{0+} \frac{s\,(s-1)}{2}\Delta^2 f_0 + ... \right] ds \qquad (24)$$

The error of interpolation of (24) is given by

$$|E_h(f)| \le \frac{h^{n+2} M_{n+1}}{(n+1)!}\int_0^1 s(s-1)(s-2)...(s-1)\,ds$$

We can obtain the trapezoidal rule (14) from (24) by using linear interpolation i.e., $f(x) \approx P_1(x) = f_0 + s\Delta f_0$. We then have

$$I_T(x) = h\int_0^1 [f_0 + s\Delta f_0]\,ds$$

$$= h\left[s f_0 + \frac{s^2}{2}\Delta f_0 \right]_0^1$$

$$= h\left[f_0 + \frac{\Delta f_0}{2} \right] = \frac{h}{2}[f_0 + f_1]$$

with the error of integration given by (15).

Similarly Simpson's rule (16) can be obtained from (24) by using quadratic interpolation i.e., $f(x) \approx P_2(x)$.

Taking $x_0 = a$, $x_1 = x_0 + h$, $x_2 = x_0 + 2h = b$, we have

$$I_s[f] = \int_a^b f(x)\,dx \approx h\int_0^2 \left[f_0 + s\Delta f_0 + \frac{s(s-1)}{2}\Delta^2 f_0 \right] ds$$

$$= h\left[2f_0 + 2\Delta f_0 + \frac{\Delta^2 f_0}{3} \right]$$

$$= \frac{h}{3}[f_n + 4f_1 + f_2].$$

The error of interpolation is given by Eqn. (18).

Example 3: Find the approximate value of $I = \int_0^1 \frac{dx}{1+x}$ using

Simpson's rule. Obtain the error bound and compare it with the actual error. Also compare the result obtained here with the one obtained in Example 1.

Solution: Here $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$ and $h = \frac{1}{2}$.

Using Simpson's rule we have

$$I_s[f] = \frac{h}{3}[f(0) + 4f(0.5) + f(1)] = \frac{1}{6}\left[1 + \frac{8}{3} + 0.5 \right] = 0.694445$$

Exact value of I =ln 2 = 0.693147.

Actual error = 0.001297. The bound for the error is given by

$$|E_s[f]| \le \frac{h^5}{90} M_4 = 0.00833, \text{ where } M_4 = \max \left| \frac{24}{(1+x)^5} \right| = 24$$

Here too the actual error is less than the given bound.

Also actual error obtained here is much less than that obtained in Example 1. You may now try the following exercise.

Ex. 2) Find an approximation to $\int_{1.1}^{1.5} e^x\,dx$, using

a) the trapezoidal rule with h = 0.4
b) Simpson's rule with h = 0.2

The Newton-Cotes formulas as derived above are generally unsuitable for use over large integration intervals. Consider for instance, an approximation to $\int_0^4 e^x\,dx$, using Simpson's rule with h = 2. Here

$$\int_0^4 e^x dx \approx \frac{2}{3}\left(e^0 + 4e^2 + e^4\right) = 56.76958.$$

Since the exact value in this case $e^4 - e^0 = 53.59815$, the error is –3.17143. This error is much larger than what we would generally regard as acceptable. However, large error is to be expected as the step length h = 2.0 is too large to make the error expression meaningful. In such cases, we would be required to use higher order formulas. An alternate approach to obtain more accurate results while using lower order methods is the use of composite integration methods, which we shall discuss in the next section.

2.3 COMPOSITE INTEGRATION

In composite integration we divide the given interval [a, b] into a number of subintervals and evaluate the integral in each of the subintervals using one of the integration rules. We shall construct composite rules of integration for trapezoidal and Simpson's methods and find the corresponding errors of integration when these composite rules are used.

Composite Trapezoidal Rule

We divide the interval [a, b] into N subintervals of length $h = \frac{(b-a)}{N}$. We denote the subintervals as

$(x_{k-1}, x_k), k = 1, 2, .., N$ where $x_0 = a, x_N = b$. Then

$$I = \int_a^b f(x)\,dx = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} f(x)\,dx \qquad (25)$$

Evaluating each of the integrals on the right hand side by trapezoidal rule, we have

$$I_T[f] = \sum_{k=1}^{N} \frac{h}{2}\left[f_{k-1} + f_k\right] = \frac{h}{2}\left[f_0 + f_N + 2\left(f_1 + f_2 + ... + f_{N-1}\right)\right] \qquad (26)$$

`

$$E_T[f] = -\frac{h^3}{12}\left[\sum_{i=1}^{N} f''(\alpha_i)\right], x_{k-1} < \alpha_i < x_k, \text{ for } k = 1...., N.$$

Now since f is a continuous function on the interval [a, b] we have as a consequence of Intermediate-value theorem

$$\sum_{i=1}^{N} f''(\alpha_i) = f''(\xi) \sum_{i=1}^{N} 1, \text{ where } a < \xi < b.$$

$$\therefore E_T[f] = \frac{-h^3}{12} f''(\xi)\, N, a < \xi < b.$$

$$= \frac{-Nh}{12} h^3 f''(\xi)$$

$$= \frac{-(b-a)h^{2.}}{12} f''(\xi).$$

If $M_2 = \max_{a < \xi < b} |f''(\xi)|$. Then

$$|E_T[f]| \le \frac{(b-a)h^2}{12} M_2 \qquad (27)$$

The error is of order h^2 and it decreases as h decreases

Composite trapezoidal rule integrates exactly polynomials of degree ≤ 1. We can try to remember the formula (26) as

$$I_T[f] = \left(\frac{h}{2}\right)[\text{first ordinate} + \text{last ordinate} + 2(\text{sum of the remaining ordinates})].$$

Composite Simpson's Rule

In using Simpson's rule of integration (17), we need three abscissas. Hence, we divide the interval (a, b) into an even number of subintervals of equal length giving an odd number of abscissas in the form $a = x_0 < x_1 < x_2 < \ldots.. < x_{2n} = b$ with $h = \frac{b-a}{2N}$ and $x_k = x_0 + kh$, k = 0, 1, 2, …, 2N. We then write

$$I = \int_a^b f(x)\,dx = \sum_{k=1}^{N} \int_{x_{2k-2}}^{x_{2k}} f(x)\,dx \qquad (28)$$

Evaluating each of the integrals on the right hand side of Eqn. (28) by the Simpson's rule, we have

) (29)

The formula (29) is known as **the composite Simpson's rule of numerical**

integration. The error in (29) is obtained from (18) by adding up the errors. Thus we get

$$E_s[f] = -\frac{h^5}{90}\left[\sum_{k=1}^{N} f^{IV}(\alpha)\right], x_{2k-2} < \alpha_k < x_{2k}$$

$$= -\frac{h^5}{90} f^{iv}(\xi) \sum_{i=1}^{N} 1, \quad a < \xi < b$$

$$= -\frac{Nh^5}{90} f^{IV}(\xi)$$

$$= -\frac{(b-a)h^4}{180} f^{IV}(\xi).$$

If $M_4 = \max_{a \le \xi \le b} |f^{IV}(\xi)|$ write using $h = \frac{(b-a)}{2N}$

$$|E_s[f]| \le \frac{(b-a)}{180} h^4 M_4 \qquad (30)$$

The error is of order h^4 and it approaches zero very fast as $h \to 0$. The rule integrates exactly polynomials of degree ≤ 3. We can remember the composite Simpson's rule as

$$I_s[f] = \left(\frac{h}{3}\right)[\text{first ordinate} + \text{last ordinate} + 2(\text{sum of even ordinates}) + 4(\text{sum of the remaining odd ordinates})]$$

We now illustrate composite trapezoidal and Simpson's rule through examples.

Example 4: Evaluate $\int_0^1 \frac{dx}{1+x}$ using

(a) Composite trapezoidal rule (b) composite Simpson's rule with 2, 4 and 8 subintervals.

Solution: We give in Table 1 the values of f (x) with h = $\frac{1}{8}$ from x = 0 to x = 1.

Table 1

x	:	0	1/8	2/8	3/8	4/8	5/8	6/8	7/8	1
f(x)	:	1 f_0	8/9 f_1	8/10 f_2	8/11 f_3	8/12 f_4	8/13 f_5	8/14 f_6	8/15 f_7	8/16 f_8

If N = 2 then h = 0.5 and the ordinates f_0, f_4 and f_8 are to be used

We get

$$I_T[f] = \frac{1}{4}[f_0 + 2f_4 + f_8] = \frac{17}{24} = 0.708333$$

$$I_s[f] = \frac{1}{6}[f_0 + 4f_1 + f_8] = \frac{25}{36} = 0.694444$$

If N = 4 then h = 0.25 and the ordinates f_0, f_2, f_4, f_6, f_8 are to be used.

We have

$$I_T[f] = \frac{1}{8}[f_0 + f_8 + 2(f_2 + f_4 + f_6)] = 0.697024$$

$$I_T[f] = \frac{1}{12}[f_0 + f_8 + 4(f_2 + f_6) + 2\,f_4] = 0.693254$$

If N = 8 then h = 1/8 and all the ordinates in Table 1 are to be used.

We obtain

$$I_T[f] = \frac{1}{16}[f_0 + f_8 + 2(f_1 + f_2 + f_3 + f_4 + f_5 + f_6 + f_7)] = 0.694122$$

$$I_s[f] = \frac{1}{24}[f_0 + f_8 + 4(f_1 + f_3 + f_5 + f_7) + 2(f_2 + f_4 + f_6)] = 0.693147$$

The exact value of the given integral correct to six decimal places is ln2 = 0.693147.

We now give the actual errors in Table 2 below.

Table 2

N	E_T (f)	E_s (f)
2	0.015186	0.001297
4	0.003877	0.000107
8	0.000975	0.000008

Note that as h decreases, the errors in both trapezoidal and Simpson's rule also decreases. Let us consider another example.

Example 5: Find the minimum number of intervals required to evaluate $\int_0^1 \frac{dx}{1+x}$ with an accuracy 10^{-6}, by using the Simpson rule.

Solution: In example 4 you may observe from Table 2 that $N \approx 8$ gives (1.E−06) accuracy. We shall now determine N from the theoretical error bound for Simpson's rule which gives 1.E – 06 accuracy. Now

$$\left|E_s[f]\right| \le \frac{(b-a)^5 M_4}{2880N^4}$$

where

$$M_4 = \max_{0<x<1} \left|f^{IV}(x)\right|$$

$$= \max_{0<x<1} \left|\frac{24}{(1+x)^5}\right| = 24$$

To obtain the required accuracy we should therefore have

$$\frac{24}{2880N^4} \le 10^{-6}, \text{ or } N^4 \ge \frac{24*10^6}{2880} = 8333.3333$$

$$\therefore N \ge 9.5$$

We find that we cannot take N = 9 since to make use of Simpson's rule we should have even number of intervals. We therefore conclude that N = 10 should be the minimum number of subintervals to obtain the accuracy 1.E – 06 (i.e., 10^{-6})

Ex.3) Evaluate $\int_0^1 \frac{dx}{1+x^2}$ by subdividing the interval (0, 1) into 6 equal parts and using.

(a) Trapezoidal rule (b) Simpson's rule. Hence find the value of π and actual errors.

Ex. 4) A function f(x) is given by the table

Ex.5) The speedometer reading of a car moving on a straight road is given by the following table Estimate the distance traveled by the car in 12 minutes using (a) Trapezoidal rule (b) Simpson's rule.

Ex.6) Evaluate $\int_{0.2}^{0.4} (\sin x - \ln x + e^x)\,dx$

using (a) Trapezoidal rule (b)

Simpson's rule taking h = 0.1. Find the actual errors.

Ex.7) Determine N so that composite trapezoidal rule gives the value of $\int_0^1 e^{-x^2}\,dx$ correct upto 3 digits after the decimal point, assuming that e^{-x^2} can be calculated accurately.

2.4 SUMMARY

In this unit, we learnt the following :

1) If a function f(x) is not known explicitly but a table of values of x is given or when it has no anti-derivative expressible in terms of elementary functions then its integral cannot be obtained by calculus methods. In such cases numerical integration methods are used to find the definite integral of f(x) using the given data.

2) The basic idea of numerical integration methods is to approximate the definite integral as a linear combination of the values of f(x) in the form

$$\int_a^b f(x)\,dx \approx \sum_{k=0}^{n} \beta_k f(x_k) \qquad \text{(See Eqn. (21))}$$

where the (n+1) distinct nodes $x_k, k = 0,1,....,n,$ with $x_0 < x_1 < x_2 < .... < x_n$ divide the interval (a, b) into n subintervals and $\beta_k, k = 0,1,....,$ n are the weights of the integration rule. The error of the integration methods is then given by

$$|E_h[f]| = \int_a^b f(x)\,dx - \sum_{k=0}^{n} \beta_k f(x_k) \qquad \text{(see Eqn. (4))}$$

3) For equispaced nodes, the integration formulas derived by using Lagrange interpolating polynomials $P_n(x)$ of degree $\leq$ n, satisfying the interpolatory conditions $P_n(x_k) = f(x_k), k = 0,1,...,$ n are known as Newton-Cotes formulas. Corresponding to n = 1 and n = 2, Newton-Cotes formulas viz., trapezoidal rule and Simpson's rule are obtained.

4) For large integration intervals, the Newton-Cotes formulas are generally unsuitable for they give large errors. Composite integration methods can be used in such cases by dividing the interval into a large number of subintervals and evaluating the integral in each of the subintervals using one of the integration rules.

2.5 SOLTIONS/ANSWERS

E1) a) $I_T[f]=\frac{h}{2}[f_0+f_1]=0.346574$

$$I_s[f]=\frac{h}{3}[f_0+4f_1+f_2]$$
$$=\frac{0.5}{3}[4\ln 1.5+\ln 2]=0.385835$$

Exact value of I = 0.386294

Actual error in $I_T[f]=0.03972$

Actual error in $I_S[f]=0.0000459$

Also

$$|E_T[f]|\le-\frac{h^3}{12}\max_{1<x<2}\left|\frac{1}{x^2}\right|=-\frac{1}{12}=-0.083334$$

$$|E_s[f]|\le-\frac{h^5}{90}\max_{1<x<2}\left|\frac{6}{x^4}\right|=-0.002083$$

b) $I_T[f]=0.023208,\ |E_T[f]|=\text{none}.$
$I_S[f]=0.032296,\ |E_S[f]|=\text{none}.$

Exact value = 0.034812.

c) $I_T[f]=0.39270,\ |E_T[f]|=0.161.$

$I_S[f]=0.34778,\ |E_S[f]|=0.00831.$

Exact value = 0.34657

E2) $I_T[f]=1.49718$

$I_S[f]=1.47754$

E3) With h = 1/6, the values of $f(x)=\frac{1}{1+x^2}$

From x = 0 to 1 are

x	:	0	1/6	2/6	3/6	4/6	5/6	1
f(x)	:	1	0.972973	0.9	0.8	0.692308	0.590167	0.5

Now

$$I_T[f]=\frac{h}{2}[f_0+f_6+2(f_1+f_2+f_3+f_4+f_5)]$$

$$=0.784241$$

$$I_S[f]=\frac{h}{3}[f_0+f_6+4(f_1+f_3+f_5)+2(f_2+f_4)]$$

$$=0.785398$$

$$\int_0^1 \frac{dx}{1+x^2}=\left[\tan^{-1}x\right]_0^1=\frac{\pi}{4}.$$ Exact $\pi = 3.141593$

Value of π from $I_T[f]=4x0.784241=3.136964$

Error in calculating π by $I_T[f]$ is $E_T[f]=0.004629$

Value of π from $I_s[f]=4\,x\,0.785398=3.141592$

Error in π by $I_s[f]$ is $E_s[f]=1.0\times10^{-6}$

E4) $I_T[f]=\left(\frac{h}{2}\right)[f_0+f_4+2(f_1+f_2+f_3)]$

$$=(1/4)[1+25+2(2.875+7+14.125)]=18.5$$

$$I_s[f]=\left(\frac{h}{3}\right)[f_0+f_4+2f_2+4(f_1+f_3)]$$

$$=(1/6)[1+25+2\,x7+(2.875+14.125)]=18$$

E5) Let $v_0=0,\ v_1=15, v_2=25, v_3=40, v_4=45, v_5=20, v_6=0.$ Then

$$I=\int_0^{12} v\,dt,$$

$$I_T[v]=\left(\frac{h}{2}\right)[v_0+v_6+2(v_1+v_2+v_3+v_4+v_5)]=290$$

$$I_s[v]=\frac{880}{30}=293.33.$$

E6) The values of f(x) = sin x = 1n x + e^x are

$$f(0.2)=3.02951, f(0.3)=2.849352, f(0.4)=2.797534$$

$$I_T[f]=\left(\frac{0.1}{2}\right)[f(0.2)+2f(0.3)+f(0.4)]=0.57629$$

$$I_S[f]=\left(\frac{0.1}{3}\right)[f(0.2)+4f(0.3)+f(0.4)]=0.574148$$

Exact value = 0.574056

$E_T = 2.234 \times 10^{-3}$

$E_s = 9.2 \times 10^{-5}$

E7) Error in composite trapezoidal rule

$$E_T[f] = -\frac{(b-a)^3}{12N^2} M_2 \,,\; M_2 = \max_{0 < x < 1} |f''(x)|.$$

Thus

$$|E_T[f]| \le \frac{1}{12N^2} \max_{0 \le x \le 1} |f''(x)|$$

$$f(x) = e^{-x^2}, f''(x) = e^{-x^2}(4x^2 - 2)$$

$$f'''(x) = e^{-x^2}\, 4x(3 - 2x^2) = 0 \text{ when or } x = 0, x = \sqrt{1.5}$$

$$\max |[f''(0), f''(1)]| = \max \left[2, 2e^{-1}\right] = 2$$

For getting the correct value upto 3 digits, we must have

$$\frac{2}{12N^2} < 10^{-03} \text{ or } N^2 > \frac{10^3}{6} = \frac{10^4}{60}$$

or

$$N > \frac{100}{\sqrt{(60)}} \approx 12.9.$$

The integer value of N =13.

UNIT 3 NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

Structure **Page Nos.**

3.0 INTRODUCTION

In the previous two units, you have seen how a complicated or tabulated function can be replaced by an approximating polynomial so that the fundamental operations of calculus viz., differentiation and integration can be performed more easily. In this unit we shall solve a differential equation, that is, we shall find the unknown function which satisfies a combination of the independent variable, dependent variable and its derivatives. In physics, engineering, chemistry and many other disciplines it has become necessary to build mathematical models to represent complicated processes. Differential equations are one of the most important mathematical tools used in modeling problems in the engineering and physical sciences. As it is not always possible to obtain the analytical solution of differential equations recourse must necessarily be made to numerical methods for solving differential equations. In this unit, we shall introduce two such methods namely, Euler's method and Taylor series method to obtain numerical solution of ordinary differential equations (ODEs). To begin with, we shall recall few basic concepts from the theory of differential equations which we shall be referring to quite often.

3.1 OBJECTIVES

After studying this unit you should be able to:

- identify the initial value problem (IVP) for the first order ordinary differential equations;
- state the difference between the single step and multi-step methods of finding solution of IVP;
- obtain the solution of the initial value problems by using single-step methods viz., Taylor series method and Euler's method.

3.2 BASIC CONCEPTS

In this section we shall state a few definitions from the theory of differential equations and define some concepts involved in the numerical solution of differential equations.

Definition: An equation involving one or more unknown function (dependent variables) and its derivatives with respect to one or more known functions (independent variables) is called **differential equation.**
For example,

$$x\frac{dy}{dx} = 2y \qquad (1)$$

$$x\frac{\partial z}{\partial x} + y\frac{\partial z}{\partial y} - z = 0 \qquad (2)$$

are differential equations.

Differential equations of the form (1), involving derivatives w.r.t. a single independent variable are called **ordinary differential equations** (ODEs) whereas, those involving derivatives w.r.t. two or more independent variables are **partial differential equations** (PDEs). Eqn. (2) is an example of PDE.

Definition: The **order** of a differential equation is the order of the highest order derivative appearing in the equation and its **degree** is the highest exponent of the highest order derivative after the equation has been rationalized i.e., after it has been expressed in the form free from radicals and any fractional power of the derivatives or negative power. For example equation

$$\left(\frac{d^3y}{dx^3}\right)^2 + 2\frac{d^2y}{dx^2} - \frac{dy}{dx} + x^2\left(\frac{dy}{dx}\right)^3 = 0 \qquad (3)$$

is of **third** order and **second** degree. Equation

$$y = x\frac{dy}{dx} + \frac{a}{dy/dx}$$

is of **first** order and **second** degree as it can be written in the form

$$y\frac{dy}{dx} = x\left(\frac{dy}{dx}\right)^2 + a \qquad (4)$$

Definition: When the dependent variable and its derivatives occur in the first degree only and not as higher powers or products, the equation is said to be **linear,** otherwise it is **nonlinear.**

Equation $\frac{d^2y}{dx^2} + y = x^2$ is a linear ODE, whereas $(x+y)^2\ \frac{dy}{dx} = 1$ is a nonlinear ODE.

Similarly, $\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x\ \partial y}\right)^2 = 0,$ is a nonlinear PDE.

In this unit we shall be concerned only with the ODEs.

The general form of a linear ODE of order n can be expressed in the form

$$L[y] \equiv a_0(t)\, y^{(n)}(t) + a_1(t)\, y^{(n-1)}(t) + \ldots\ldots + a_{n-1}(t)\, y'(t) + a_n(t)\, y(t) = r(t) \qquad (5)$$

where r(t), $a_i(t)$, i = 0,1, 2, ……., n are known functions of t and

$$L = a_0(t)\frac{d^n}{dt^n} + a_1(t)\frac{d^{n-1}}{dt^{n-1}} + \ldots\ldots\ldots\ldots + a_{n-1}(t)\frac{d}{dt} + a_n(t),$$

is the linear differential operator. The general nonlinear ODE of order n can be written as

$$F(t, y, y', y'', \ldots\ldots\ldots\ldots, y^{(n)}) = 0 \qquad (6)$$

or, $y^{(n)} = f(t, y, y', y'', \ldots\ldots\ldots\ldots, y^{(n-1)})$. (7)

Eqn. (7) is called a **canonical representation** of Eqn. (6). In such a form, the highest order derivatives is expressed in terms of lower order derivatives and the independent variable.

The **general solution** of an nth order ODE contains n arbitrary constants. In order to determine these arbitrary constants, we require n conditions. If all these conditions are given at one point, then these conditions are known as **initial conditions** and the differential equation together with the initial conditions is called an **initial value problem** (IVP). The nth order IVP alongwith associates initial conditions can be written as

$$y^{(n)}(t) = f(t, y, y', y'', \ldots\ldots\ldots, y^{(n-1)})$$
$$y^{(p)}(t_0) = y_0^{(p)},\ p = 0, 1, 2, \ldots\ldots\ldots, n-1. \qquad (8)$$

We are required to find the solution y(t) for $t > t_o$

If the n conditions are prescribed at more than one point then these conditions are known as **boundary conditions.** These conditions are prescribed usually at two points, say t_0 and t_a and we are required to find the solution y(t) between $t_o < t < t_a$ .The differential equation together with the boundary conditions is then known as a **boundary value problem** (BVP).

As may be seen below, the nth order IVP (8) is equivalent to solving following system of n first order equations:

Setting $y = y_1$,

Similarly setting $y'_{i-1} = y_i$, we may write

$$y' = y'_1 = y_2 \qquad y_1(t_0) = y_0$$
$$y'_2 = y_3 \qquad y_2(t_0) = y'_0$$
$$\ldots\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots$$
$$y'_{n-1} = y_n \qquad y_{n-1}(t_0) = y_0^{(n-2)}$$

$$y'_n = f(t, y_1, y_2, \ldots\ldots\ldots, y_n) \qquad y_n(t_0) = y_0^{(n-1)};$$

In vector notation, this system can be written as a single equation as

$$\frac{dy}{dx} = f(t, y), \qquad y(t_0) = \alpha \qquad (9)$$

where $y = (y_1, y_2 \ldots\ldots\ldots, y_n)^T$, $f(t,y) = (y_2, y_3, \ldots\ldots\ldots, f(t, y_1, \ldots\ldots, y_n))^T$

$$\alpha\ (y_0, y'_0, \ldots\ldots\ldots, y_0^{(n-1)})^T.$$

Hence, it is sufficient to study numerical methods for the solution of the first order IVP.

$$y' = f(t, y),\ y(t_0) = y_0 \qquad (10)$$

The vector form of these methods can then be used to solve Eqn. (9). Before attempting to obtain numerical solutions to Eqn, (10), we must make sure that the

problem has a unique solution. The following theorem ensures the existence and uniqueness of the solution to IVP (10).

Theorem 1: If f(t, y) satisfies the conditions

i) f(t, y) is a real function
ii) f(t, y) is bounded and continuous for $t \in [t_0, b]$, $y \in]-\infty, \infty[$
iii) there exists a constant L such that for any $t \in [t_0, b]$ and for any two numbers y_1 and y_2

$$| f(t, y_1) - f(t, y_2) | \leq L | y_1 - y_2 |$$

then for any y_0, the IVP (10) has a unique solution. This condition is called the **Lipschitz condition** and L is called the **Lipschitz constant.**

We assume the existence and uniqueness of the solution and also that f(t,y) has continuous partial derivatives w.r.t. t and y of as high order as we desire.

Let us assume that $[t_0, b]$ be an interval over which the solution of the IVP (10) is required. If we subdivide the interval $[t_0, b]$ into n subintervals using a step size $h = \left[\frac{t_n - t_o}{n}\right]$, where $t_n = b$, we obtain the **mesh points** or **grid points** $t_0, t_1, t_2, \ldots, t_n$ as shown in Fig. 1.

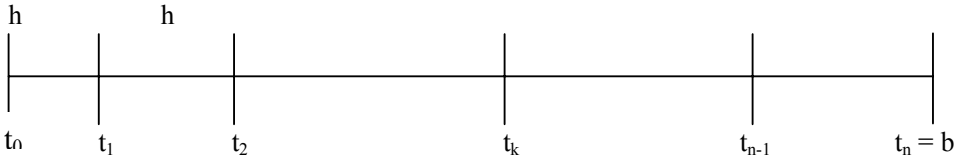


Fig. 1

We can then write $t_k = t_0 + kh$, k = 0, 1, ….., n. A numerical method for the solution of the IVP (10), will produce approximate values y_k at the grid points t_k in a step by step manner i.e. values of $y_1, y_2, \ldots$ etc in unit order.

Remember that the approximate values y_k may contain the truncation and round-off errors. We shall now discuss the construction of numerical methods and related basic concepts with reference to a simple ODE.

$$\frac{dy}{dt} = \lambda y, \quad t \varepsilon [a\ b] \qquad (11)$$

$$y(t_0) = y_0,$$

where λ in a constant.

Let the domain [a, b] be subdivided into N intervals and

let the grid points be defined by,

$$t_j = t_0 + jh, \quad j = 0, 1, \ldots\ldots\ldots\ldots, N$$

where $t_0 = a$ and $t_N = t_0 + Nh = b$.

Separating the variables and integrating, we find that the exact solution of Eqn. (11) is

$$y(t) = y(t_0)\, e^{\lambda(t-t_0)}. \qquad (12)$$
$$y(t_0)\, e^{\lambda(t-t0)}$$

In order to obtain a relation between the solutions at two successive permits, $t = t_n$ and t_{n+1} in Eqn. (12), we use,

$$y(t_n) = y(t_0)\ e^{\lambda(t_n - t_0)}$$

and

$$y(t_{n+1}) = y(t_0)\ e^{\lambda(t_{n+1} - t_0)}\ .$$

Dividing we get

$$\frac{y(t_{n+1})}{y(t_n)} = \frac{e^{\lambda t_{n+1}}}{e^{\lambda t_n}} = e^{\lambda(t_{n+1} - t_n)}.$$

Hence we have,

$$y(t_{n+1}) = e^{\lambda h}\, y\,(t_n),\ n = 0, 1, \ldots.., N-1. \tag{13}$$

Eqn. (13) gives the required relation between $y\,(t_n)$ and $y(t_{n+1})$.

Setting $n = 0, 1, 2, \ldots\ldots,$ N-1, successively, we can find $y(t_1)$, $y(t_2)$,…….., $y(t_N)$ from the given value $y(t_0)$.

An approximate method or a numerical method can be obtained by approximating $e^{\lambda h}$ in Eqn. (13). For example, we may use the following polynomial approximations,

$$e^{\lambda h} = 1 + \lambda h + 0\,(|\lambda h|^2) \tag{14}$$

$$e^{\lambda h} = 1 + \lambda h + \frac{\lambda^2 h^2}{2} + 0\,(|\lambda h|^3) \tag{15}$$

$$e^{\lambda h} = 1 + \lambda h + \frac{\lambda^2 h^2}{2} + \frac{\lambda^3 h^3}{6} + 0\,(|\lambda h|^4) \tag{16}$$

and so on.

Let us retain (p + 1) terms in the expansion of $e^{\lambda h}$ and denote the approximation to $e^{\lambda h}$ by $E(\lambda h)$. The numerical method for obtaining the approximate values y_n of the exact solution $y(t_n)$ can then be written as

$$y_{n+1} = E(\lambda h)\, y_n,\ n = 0, 1, \ldots\ldots\ldots, N-1 \tag{17}$$

The truncation error (TE) of the method is defined by

$$TE = y\,(t_{n+1}) - y_{n+1}.$$

Since (p +1) terms are retained in the expansion of $e^{\lambda h}$, we have

$$TE = \left(1 + \lambda h + \ldots\ldots\ldots + \frac{(\lambda h)^p}{p!} + \frac{(\lambda h)^{p+1}}{(p+1)!} e^{\theta \lambda h}\right) - \left(1 + \lambda h + \ldots\ldots + \frac{(\lambda h)^p}{p!}\right)$$

$$= \frac{(\lambda h)^{p+1}}{(p+1)!} e^{\theta \lambda h}, \qquad 0 < \theta < 1$$

The TE is of order p+1 and the numerical method is called of order p.

The concept of stability is very important in a numerical method.

We say that a numerical method is **stable** if the error at any stage, i.e. $y_n - y(t_n) = \epsilon_n$ remains bounded as $n \to \infty$. Let us examine the stability of the numerical method (17). Putting $y_{n+1} = y(t_{n+1}) + \epsilon_{n+1}$ and $y_n = y(t_n) + \epsilon_n$ in Eqn. (17), we get

$$y(t_{n+1}) + \epsilon_{n+1} = E(\lambda h)\,[y(t_n) + \epsilon_n]$$

$$\epsilon_{n+1} = E(\lambda h)\,[y(t_n) + \epsilon_n] - y(t_{n+1})$$

which on using eqn. (13) becomes

$$\epsilon_{n+1} = E(\lambda h)\,[y(t_n) + \epsilon_n] - e^{\lambda h} y(t_n)$$

$$\therefore \quad \epsilon_{n+1} = [E(\lambda h) - e^{\lambda h}]\, y(t_n) + E(\lambda h)\,\epsilon_n \qquad (18)$$

We **note** from Eqn. (18) that the error at t_{n+1} consists of two parts. The first part $E(\lambda h) - e^{\lambda h}$ is the **local truncation error** and can be made as small as we like by suitably determining $E|\lambda h|$. The second part $|E(\lambda h)|\ \epsilon_n$ is the **propagation error** from the previous step t_n to t_{n+1} and will not grow if $|E(\lambda h)| < 1$. If $|E(\lambda h)| < 1$, then as $n \to \infty$ the propagation error tends to zero and method is said to be absolutely stable. Further, a numerical method is said to be **relatively stable** if $|E(\lambda h)| \le e^{\lambda h}$, $\lambda > 0$.

The polynomial approximations (14), (15) and (16) always give relatively stable methods. Let us now find when the methods $y_{n+1} = E(\lambda h)\, y_n$ are absolutely stable where $E(\lambda h)$ is given by (14) (15) or (16).

These methods are given by

First order: $y_{n+1} = (1 + \lambda h)\, y_n$

Second order: $y_{n+1} = \left(1 + \lambda h + \frac{\lambda^2 h^2}{2}\right) y_n$ and

Third order: $y_{n+1} = \left(1 + \lambda h + \frac{\lambda^2 h^2}{2} + \frac{\lambda^3 h^3}{6}\right) y_n$

Let us examine the conditions for absolute stability in various methods:

First order: $|1 + \lambda h| \le 1$
or $-1 \le 1 + \lambda h \le 1$
or $-2 \le \lambda h \le 0$

Second order: $\left|1 + \lambda h + \frac{\lambda^2 h^2}{2}\right| \le 1$

or $-1 \le 1 + \lambda h + \frac{\lambda^2 h^2}{2} \le 1$

The right inequality gives

$$\lambda h \left(1 + \frac{\lambda h}{2}\right) \le 0$$

i.e., $\lambda h \le 0$ and $1 + \frac{\lambda h}{2} \ge 0$.

The second condition gives $-2 \le \lambda h$. Hence the right inequality gives $-2 \le \lambda h \le 0$.

The left inequality gives

$$2 + \lambda h + \frac{\lambda^2 h^2}{2} \geq 0.$$

For $-2 \leq \lambda h \leq 0$, this equation is always satisfied. Hence the stability condition is $-2 \leq \lambda h \leq 0$.

Third order: $\left| 1 + \lambda h + \frac{\lambda^2 h^2}{2} + \frac{\lambda^3 h^3}{6} \right| \leq 1$

Using the right and left inequalities, we get

$$-2.5 \leq \lambda h \leq 0.$$

These intervals for λh are known as **stability intervals.**

Numerical methods for finding the solution of IVP given by Eqn. (10) may be broadly classified as,

i) Singlestep methods
ii) Multistep methods

Singlestep methods enable us to find y_{n+1}, an approximation to $y(t_{n+1})$, in terms of y_n, and y'_n.

Multistep methods enable us to find y_{n+1}, an approximation to $y(t_{n+1})$, in terms of y_i, y'_i , i = n, n-1, ……n-m+1 i.e. values of y and y′ at previous m points. Such methods are called m-step multistep methods.

In this course we shall be discussing about the singlestep methods only.

A singlestep method for the solution of the IVP

$y' = f(t, y), \quad y(t_0) = y_0, t \in (t_0, b)$

is a recurrence relation of the form

$$y_{n+1} = y_n + h \phi (t_n, y_n, h) \qquad (19)$$

where $\phi (t_n, y_n, h)$ is known as the **increment function.**

If y_{n+1} can be determined from Eqn. (19) by evaluating the right hand side, then the singlestep method is known as an **explicit method,** otherwise it is known as an **implicit method.** The local truncation error of the method (19) is defined by

$$TE = y (t_{n+1}) - y(t_n) - h \phi (t_n, y_n, h) . \qquad (20)$$

The largest integer p such that

$$\left| h^{-1} TE \right| = O (h^p) \qquad (21)$$

is called the **order** of the singlestep method.

Let us now take up an example to understand how the singlestep method works.

Example 1: find the solution of the IVP $y' = \lambda y$, $y(0) = 1$ in $0 < t \leq 0.5$, using the first order method

$$y_{n+1} = (1 + \lambda h) y_n \text{ with } h = 0.1 \text{ and } \lambda = \pm 1.$$

Solution: Here the number of intervals are N = $\frac{0.5}{h} = \frac{0.5}{0.1} = 5$

We have $y_0 = 1$

$y_1 = (1 + \lambda h)\, y_0 = (1 + \lambda h) = (1 + 0.1\lambda)$

$y_2 = (1 + \lambda h)\, y_1 = (1 + \lambda h)^2 = (1 + 0.1\lambda)^2$

$y_5 = (1 + \lambda h)^5 = (1 + 0.1\lambda)^5$

The exact solution is $y(t) = e^{\lambda t}$.

We now give in Table 1 the values of y_n for $\lambda = \pm 1$ together with exact values.

Table 1

Solution of $y' = \lambda y$, $y(0) = 1$, $0 \le t \le 0.5$ with h = 01.				
$\lambda = 1$			$\lambda = -1$	
t	First Order method	Exact Solution	First Order method	Exact Solution
0	1	1	1	1
0.1	1.1	1.10517	0.9	0.90484
0.2	1.21000	1.22140	0.81	0.81873
0.3	1.33100	1.34986	0.729	0.74082
0.4	1.46410	1.49182	0.6561	0.67032
0.5	1.61051	1.64872	0.59049	0.60653

Similarly you can obtain the solution using the second order method and compare the results obtained in the two cases.

Ex 1) Find the solution of the IVP

$y' = \lambda y$, $y(0) = 1$

in $0 \le t \le 0.5$ using the second order method

$y_{n+1} = \left(1 + \lambda h + \frac{\lambda^2 h^2}{2}\right) y_n$ with h = 0.1 and $\lambda = 1$.

We are now prepared to consider numerical methods for integrating differential equations. The first method we discuss is the Taylor series method. It is not strictly a numerical method, but it is the most fundamental method to which every numerical method must compare.

3.3 TAYLOR SERIES METHOD

Let us consider the IVP given by Eqn. (10), i.e.,

$$y' = f(t,y),\ y(t_0) = y_0, \qquad t \in [t_0, b]$$

The function f may be linear or nonlinear, but we assume that f is sufficiently differentiable w.r.t. both t and y.

The Taylor series expansion of y(t) about any point t_k is given by

$$y(t) = y(t_k) + (t-t_k)\, y'(t_k) + \frac{(t-t_k)^2}{2!} y''(t_k) + \ldots\ldots + \frac{(t-t_k)^p}{p!} y^{(p)}(t_k) + \ldots\ldots \qquad (22)$$

Substituting $t = t_{k+1}$ in Eqn. (22), we have

$$y(t_{k+1}) = y(t_k) + hy'(t_k) + \frac{h^2 y''(t_k)}{2!} + \ldots\ldots + \frac{h^p y^{(p)}(t_k)}{p!} + \ldots\ldots \quad (23)$$

where $t_{k+1} = t_k + h$. Neglecting the terms of order h^{p+1} and higher order terms, we have the approximation

$$y_{k+1} = y_k + hy'_k + \frac{h^2}{2!} y''_k + \ldots\ldots + \frac{h^p}{p!} y_k^{(p)} \quad (24)$$

$$= y_k + h\,\phi\,(t_k, y_k, h)$$

where $\phi\,(t_k, y_k, h) = y'_k + \frac{h}{2!} y''_k + \ldots\ldots + \frac{h^{p-1}}{p!} y_k^{(p)}$

This is called the Taylor Series method of order p. The truncation error of the method is given by

$$TE = y\,(t_{k+1}) - y(t_k) - h\phi\,(t_k, y(t_k), h) \quad (25)$$

$$= \frac{h^{p+1}}{(p+1)!} y^{(p+1)}\,(t_k + \theta h), 0 < \theta < 1$$

when p = 1, we get from Eqn. (24)

$$y_{k+1} = y_k + hy'_k \quad (26)$$

which is the Taylor series method of order one.

To apply (24), we must know $y(t_k)$, $y'(t_k)$, $y''(t_k)$……., $y^{(p)}\,(t_k)$.

However, $y(t_k)$ is known to us and if f is sufficiently differentiable, then higher order derivatives can be obtained by calculating the total derivative of the given differential equation w.r.t. t, keeping in mind that y is itself a function of t. Thus we obtain for the first few derivatives as:

$$y' = f(t,y)$$
$$y'' = f_t + f\,f_y$$
$$y''' = f_{tt} + 2\,f\,f_{ty} + f^2 f_{yy} + f_y\,(f_t + ff_y) \text{ etc.}$$

where $f_t = \partial f / \partial t, f_{tt} = \partial^2 f / \partial t^2$, $f_{ty} = \frac{\partial^2 f}{\partial t 2 y}$ etc.

The number of terms to be included in the method depends on the accuracy requirements.

Let p = 2. Then the Taylor Series method of $O(h^2)$ is

$$y_{k+1} = y_k + hy'_k + \frac{h^2 y''_k}{2} \quad (27)$$

with the TE = $\frac{h^3}{6} y'''(\alpha), t_n < \alpha < t_{n+1}$

The Taylor series method of $O(h^3)$, (p = 3) is

$$y_{k+1} = y_k + hy'_k + \frac{h^2 y''_k}{2} + \frac{h^3}{6} y'''_k \quad (28)$$

with the TE = $\frac{h^4}{24} y^{(IV)}\,(\alpha),\ t_k \le \alpha \le t_{k+1}$

Let us consider the following examples.

Example 2: Using the third order Taylor series method find the solution of the differential equation.

$xy' = x-y$, $y = 2$ at $x = 2$, taking $h = 1$.

Solution: We have the derivatives and their values at x=2, y=2 as follows:

$y(2) = 2$ is given. Further, $xy' = x-y$ can be written as

$$y' = 1 - \frac{y}{x}$$

$$y'(2) = 1 - \frac{2}{2} = 0.$$

Differentiating w.r.t. x, we get

$$y'' = 0 - \frac{y'}{x} + \frac{y}{x^2}$$

$$y''(2) = - \frac{0}{2} + \frac{2}{4} = \frac{1}{2}.$$

$$y' = 1 - \frac{y}{k} \qquad = 1 - \frac{2}{2} = 1 - 1$$

Similarly,

$$y''' = -\frac{y''}{x} + \frac{2y'}{x^2} - \frac{2y}{x^3}, \qquad y'''(2) = -3/4$$

$$y'''(2) = -\frac{1}{4} + \frac{2 \times 0}{4} - \frac{2 \times 2}{8}$$

$$= \frac{3}{4}$$

Using taylor series method of $O(h^3)$ given by Eqn. (28), we obtain

$$y(2 + .1) = y(2) + 01 \times y'(2) + \frac{(.1)^2}{2} y''(2) + \frac{(.1)^3}{6} y'''(2)$$

or

$$y(2.1) = 2 + .1 \times 0 + .005 \times .5 + .001 \times \frac{1}{6} \times (-.75)$$

$$= 2 + 0.0025 - 0.000125 = 2.002375.$$

Example 3: Solve the equation $x^2y' = 1 - xy - x^2y^2$, $y(1) = -1$ from x=1 to x=2 by using Taylor series method of $O(h^2)$ with h = 1/3 and ¼ and find the actual error at x=2 if the exact solution is $y = -1/x$.

Solution: From the given equation, we have $y' = \frac{1}{x^2} - \frac{y}{x} - y^2$

Differentiating it w.r.t. x, we get

$$y'' = \frac{-2}{x^3} - \frac{y'}{x} + \frac{y}{x^2} - 2yy'$$

Using the second order method (27),

$$y_{k+1} = y_k + hy'_k + \frac{h^2}{2} y''k$$

We have the following results

$h = \frac{1}{3}$, $y(1) = -1$, $y'(1) = 1$, $y''(1) = -2$.

$x_1 = \frac{4}{3}$, $y(x_1) = -0.7778$, $y'(x_1) = 0.5409$, $y''(x_1) = -0.8455$.

$x_2 = \frac{5}{3}$, $y(x_2) = -0.6445$, $y'(x_2) = 0.3313$, $y''(x_2) = -0.4358$.

$x_3 = 2$, $y(x_3) = -0.5583$ $= y(2)$

$h = \frac{1}{4}$

$x_1 = \frac{5}{4}$, $y(x_1) = -0.8125$, $y'(x_1) = 0.6298$, $y''(x_1) = -1.0244$.

$x_2 = \frac{3}{2}$, $y(x_2) = -0.6871$, $y'(x_2) = 0.4304$, $y''(x_2) = -0.5934$.

$x_3 = \frac{7}{4}$, $y(x_3) = -0.5980$, $y'(x_3) = 0.3106$, $y''(x_3) = -0.3745$.

$x_4 = 2$, $y(x_4) = -0.5321$ $= y(2)$

Since the exact value is $y(2) = -0.5$, we have the actual errors as

$e_1 = 0.0583$ with $h = \frac{1}{3}$

$e_2 = 0.0321$ with $h = \frac{1}{4}$

Note that error is small when the step size h is small.
You may now try the following exercise.

Write the Taylor series method of order four and solve the IVPs E2) and E3).

E2) $y' = x - y^2$, $y(0) = 1$. Find $y(0.1)$ taking $h = 0.1$.

E3) $y' = x^2 + y^2$, $y(0) = 0.5$. Find $y(0.4)$ taking $h = 0.2$.

E4) Using second order Taylor series method solve the IVP

$y' = 3x + \frac{y}{2}$, $y(0) = 1$. Find $y(0.6)$ taking $h = 0.2$ and $h = 0.1$.

Find the actual error at $x = 0.6$ if the exact solution is $y = -6x - 12$.

Notice that though the Taylor series method of order p give us results of desired accuracy in a few number of steps, it requires evaluation of the higher order derivatives and becomes tedious to apply if the various derivatives are complicated. Also, it is difficult to determine the error in such cases. We now consider a method, called **Euler's method** which can be regarded as Taylor series method of order one and avoids these difficulties.

3.4 EULER'S METHOD

Let the given IVP be

$$y' = f(t, y), y(t_0) = y_0.$$

Let $[t_0, b]$ be the interval over which the solution of the given IVP is to be determined. Let h be the steplength. Then the nodal points are defined by $t_k = t_0 + kh$, $k = 0, 1, 2, \ldots\ldots\ldots\ldots, N$ with $t_N = t_0 + Nh = b$.

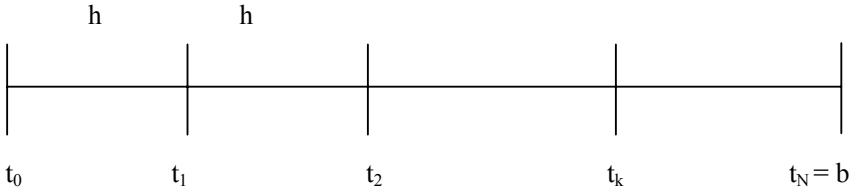


Fig. 1

The exact solution y(t) at $t = t_{k+1}$ can be written by Taylor series as

$$y(t_k + h) = y(t_k) + hy'(t_k) + \left(\frac{h^2}{2}\right) y''(t_k) + \ldots\ldots \quad (29)$$

Neglecting the term of $O(h^2)$ and higher order terms, we get

$$y_{k+1} = y_k + hy'_k \quad (30)$$

$$\text{with TE} = \left(\frac{h^2}{2}\right) y''(\alpha), t_k < \alpha < t_{k+1} \quad (31)$$

From the given IVP, $y'(t_k) = f(t_k, y_k) = f_k$

We can rewrite Eqn. (30) as

$$y_{k+1} = y_k + h f_k$$

$$\text{for } k = 0, 1, \ldots\ldots\ldots, N-1. \quad (32)$$

Eqn. (32) is known as the Euler's method and it calculates successively the solution at the nodal points t_k, $k = 1, \ldots\ldots\ldots, N$.

Since the truncation error (31) is of order h^2, Euler's method is of first order. It is also called an O(h) method.

Let us now see the geometrical representation of the Euler's method.

Geometrical Interpretation

Let y(t) be the solution of the given IVP. Integrating $\frac{dy}{dt} = (t, y)$ from t_k to t_{k+1}, we get

$$\int_{t_k}^{t_{k+1}} \frac{dy}{dt} dt = \int_{t_k}^{t_{k+1}} f(t, y)\, dt = y(t_{k+1}) - y(t_k). \quad (33)$$

We know that geometrically f(t, y) represents the slope of the curve y(t). Let us approximate the slope of the curve between t_k and t_{k+1} by the slope at t_k only. If we approximate $y(t_{k+1})$ and $y(t_k)$ by y_{k+1} and y_k respectively, then we have

$$y_{k+1} - y_k = f(t_k, y_k) \int_{t_k}^{t_{k+1}} dt \quad (34)$$

$$= (t_{k+1} - t_k)\, f(t_k, y_k)$$

$$= hf(t_k, y_k)$$

$\therefore\ y_{k+1} = y_k + hf(t_k, y_k),\ k = 0, 1, 2, \ldots, N-1.$

Thus in Euler's method the actual curve is approximated by a sequence of the segments and the area under the curve is approximated by the area of the quadrilateral.

Let us now consider the following examples.

Example 4: Use Euler method to find the solution of $y' = t + y$, given $y(0) = 1$. Find the solution on [0, 0.8] with $h = 0.2$.

Solution: We have

$$y_{n+1} = y_n + hf_n$$

$$y(0.2) \approx y_1 = y_0 + (0.2)\, f_0$$
$$= 1 + (0.2)\,[0 + 1] = 1.2$$

$$y(0.4) \approx y_2 = y_1 + (0.2)\, f_1$$
$$= 1.2 + (0.2)\,[0.2 + 1.2]$$
$$= 1.48$$

$$y(0.6) \approx y_3 = y_2 + (0.2)\, f_2$$
$$= 1.48 + (0.2)\,[0.4 + 1.48]$$
$$= 1.856$$

$$y(0.8) \approx y_4 = y_3 + (0.2)\, f_3$$
$$= 1.856 + (0.2)\,[0.6 + 1.856]$$
$$= 2.3472$$

Example 5: Solve the differential equation $y' = t+y$, $y(0) = 1$. $t \in [0, 1]$ by Euler's method using $h = 0.1$. If the exact value is $y(1) = 3.436564$, find the exact error.

Solution: Euler's method is

$$y_{n+1} = y_n + hy'_n$$

For the given problem, we have

$$y_{n+1} = y_n + h\,[t_n + y_n]$$
$$(1 + h)\, y_n + ht_n.$$

$h = 0.1$, $y(0) = 1$,
$y_1 = y_0 = (1+0.1) + (0.1)\,(0) = 1.1$
$y_2 = (1.1)\,(1.1) + (0.1)\,(0.1) = 1.22$, $y_3 = 1.362$
$y_4 = 1.5282$, $y_5 = 1.72102$, $y_6 = 1.943122$,
$y_7 = 2.197434$, $y_8 = 2.487178$, $y_9 = 2.815895$
$y_{10} = 3.187485 \approx y(1)$
exact error = $3.436564 - 3.187485 = .249079$

Example 6: Using the Euler's method tabulate the solution of the IVP

$$y' = -2ty^2,\ y(0) = 1$$

in the interval [0,1] taking $h = 0.2, 0.1$.

Solution: Euler's method gives

$$y_{k+1} = y_k + h\, f_k \text{ where } t_k = -2t_k\ y_k^2$$
$$= y_k - 2h\, t_k\ y_k^2.$$

Starting with $t_0 = 0$, $y_0 = 1$, we obtain the following table of values for $h = 0.2$.

Table 2: h = 0.2

T	y(t)
0.2	0.92
0.4	
0.6	0.78458
0.8	0.63684
1.0	0.50706

Thus y(1.0) = 0.50706 with h = 0.2
Similarly, starting with $t_0 = 0$, $y_0 = 1$, we obtain the following table of values for h = 0.1.

Table 3: h = 0.1

t	y(t)	T	y(t)
0.1	1.0	0.6	0.75715
0.2	0.98	0.7	0.68835
0.3	0.94158	0.8	0.62202
0.4	0.88839	0.9	0.56011
0.5	0.82525	1.0	0.50364

y(1.0) = 0.50364 with h = 0.1.

Remark: Since the Euler's method is of O(h), it requires h to be very small to obtain the desired accuracy. Hence, very often, the number of steps to be carried out becomes very large. In such cases, we need higher order methods to obtain the required accuracy in a limited number of steps.

Euler's method constructs $y_k \approx y(t_k)$ for each k = 1, 2, …., N,
where

$$y_{k+1} = y_k + hf(t_k, y_k).$$

This equation is called the **difference equation** associated with Euler's method. A difference equation of order N is a relation involving $y_n, y_{n+1}, \ldots\ldots, y_{n+N}$. Some simple difference equations are

$$\left.\begin{aligned} y_{n+1} - y_n &= 1 \\ y_{n+1} - y_n &= n \\ y_{n+1} - (n+1)\, y_n &= 0 \end{aligned}\right] \qquad (35)$$

where n is an integer.

A difference equation is said to be **linear** if the unknown function y_{n+k} (k = 0, 1, ……., N) appear linearly in the difference equation. The general form of a linear non-homogeneous difference equation of order N is

$$y_{n+N} + a_{N-1}\, y_{n+N-1} + \ldots\ldots\ldots + a_0\, y_n = b \qquad (36)$$

where the coefficients $a_{N-1}, a_{N-2}, \ldots\ldots\ldots\ldots, a_0$ and b may be functions of n but not of y. All the Eqns. (35) are linear. It is easy to solve the difference Eqn. (36), when the coefficients are constant or a linear or a quadratic function of n.
The general solution of Eqn. (36) can be written in the form

$$y_n = y_n(c) + y_n^{(p)}$$

where y_n (c) is the complementary solution of the homogenous equation associated with Eqn. (36) and $y_n(p)$ is a particular solution of Eqn. (36). To obtain the complementary solution of the homogeneous equations, we start with a solution in the

form $y_n = \beta^n$ and substitute it in the given equation. This gives us a polynomial of degree N. We assume that its roots $\beta_1, \beta_2, \ldots\ldots\ldots, \beta_N$ are all real and distinct.

Therefore, the general solution of the given problem is

$$y_k = C\,(1+3h)^k - \frac{5}{3}$$

Using the condition y(0) = 1, we obtain C = 8/3.

Thus

$$y_k = \frac{8}{3}(1+3h)^k - \frac{5}{3}.$$

Eqn. (41) gives the formula for obtaining $y_k \; \forall$ k.

$$y_6 = y(0.6) = \frac{8}{3}(1+3\times 0.1)^6 - \frac{5}{3}$$

$$= 11.204824.$$

Now Euler's method is

$$y_{k+1} = (1 + 3h)\, y_k + 5h$$

and we get for h = 0.1

$$y_1 = 1.8,\ y_2 = 2.84,\ y_3 = 4.192,\ y_4 = 5.9496,\ y_5 = 8.23448,\ y_6 = 11.204824.$$

You may now try the following exercises

Solve the following IVPs using Euler's method

E5) $y' = 1 - 2xy$, $y(0.2) = 0.1948$. Find y(0.4) with h = 0.2

E6) $y' = \dfrac{1}{x^2 - 4y'}$, $y(4) = 4$. Find y(4.1) taking h = 0.1

E7) $y' = \dfrac{y - x}{y + x'}$, $y(0) = 1$. Find y(0.1) with h = 0.1

E8) $y' = 1 + y^2$, $y(0) = 1$. Find y(0.6) taking h = 0.2 and h = 0.1

E9) Use Euler's method to solve numerically the initial value problem $y' = t + y$, $y(0) = 1$ with h = 0.2, 0.1 and 0.05 in the interval [0.06].

3.5 SUMMARY

In this unit, we have covered the following

1) $y' = f(t, y)$, $y(t_0) = y_0$, $t \in [t_0, b]$ is the initial value problem (IVP) for the first order ordinary differential equation.
2) Singlestep methods for finding the solution of IVP enables us to find y_{n+1}, if y_n, y'_n and h are known.
3) Multistep method for IVP enables as to find y_{n+1}, if y_i, y'_i, i = n, n-1, ….., n-m+1 and h are known and are called m-step multistep methods.

4) Taylor series method of order p for the solution of the IVP is given by

$y_{k+1} = y_k + h\,\phi\,[t_k, y_k, h]$

where $\phi\,[tk, yk, h] = y'k + \frac{h}{2!}y''_k + \ldots\ldots\ldots + \frac{h^{p-1}}{p!}y_k^{(p)}$ and $t_k = t_0 + kh, k = 0, 1, 2,$ ……..N-1, $t_N = b$. The error of approximation is given by

$TE = \frac{h^{p+1}}{(p+1)!}y^{(p+1)}(t_k + \theta h),\ 0 < \theta < 1.$

5) Euler's method is the Taylor series method of order one. The steps involved in solving the IVP given by (10) by Euler's method are as follows:

Step 1: Evaluate $f(t_0, y_0)$
Step 2: Find $y_1 = y_0 + h\,f(t_0, y_0)$
Step 3: If $t_0 < b$, change t_0 to $t_0 + h$ and y_0 to y_1 and repeat steps 1 and 2
Step 4 : If $t_0 = b$, write the value of y_1.

3.6 SOLUTIONS/ANSWERS

E1) We have $y_0 = 1, \lambda = 1, h = 0.1$

$$y_1 = \left(1 + 0.1 + \frac{(0.1)^2}{2}\right)$$

$y_2 = (1.105)^2$

$y_5 = (1.105)^5$

Table giving the values of y_n together with exact values is

Table 4

t	Second order method	Exact solution
0	1	1
0.1	1.105	1.10517
0.2	1.22103	1.22140
0.3	1.34923	1.34986
0.4	1.49090	1.49182
0.5	1.64745	1.64872

E2) Taylor series method of $0(h^4)$ to solve $y' = x\text{-}y^2$, $y(0) = 1$ is

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2}y''_n + \frac{h^3}{6}y'''_n + \frac{h^4}{24}y_n^{iv}$$

$y' = x - y^2$	$y'(0) = -1$
$y'' = 1\text{-}2xy'$	$y''(0) = 3$
$y''' = -2xy'' - 2(y')^2$	$y'''(0) = -8$
$y^{iv} = -2yy'' - 6y'\,y''$	$y^{iv}(0) = 34$

Substituting

$$y(0.1) = 1 - (0.1)(-1) + \frac{(0.1)^2}{2}(3) + \frac{(0.1)^3}{6}(-8) + \frac{(0.1)^4}{24}(34)$$
$$= 0.9138083$$

E3) Taylor series method

$y' = x^2 + y^2$, $y(0) = 0.5$, $y'(0) = 0.25$, $y'(0.2) = 0.35175$

$y'' = 2x + 2yy'$ $y''(0) = 0.25,$ $y''(0.2) = 0.79280$
$y''' = 2 + 2yy'' + 2(y')^2$ $y'''(0) = 2.375,$ $y'''(0.2) = 3.13278$
$y^{iv} = 2yy''' + 6y'y''$ $y^{iv}(0) = 2.75,$ $y^{iv}(0.2) = 5.17158$

E4) Second order Taylor's method is

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2} y''_n.$$

h = 0.2

$y(0) = 1.$ $y'(0) = 0.5,$ $y''(0) = 3.25$
$y(0.2) = 1.165,$ $y'(0.2) = 1.1825,$ $y''(0.2) = 3.59125$
$y(0.4) = 1.47333,$ $y'(0.4) = L93667,$ $y''(0.4) = 3.96833$
$y(0.6) = 1.94003$

h = 0.1

$y(0.1) = 1.06625,$ $y'(0.1) = 0.83313,$ $y''(0.1) = 3.41656$
$y(0.2) = 1.16665,$ $y'(0.2) = 1.18332,$ $y''(0.2) = 3.59167$
$y(0.3) = 1.46457,$ $y'(0.3) = 1.63228,$ $y''(0.3) = 3.81614$
$y(0.4) = 1.64688,$ $y''(0.4) = 2.02344,$ $y''(0.4) = 4.01172$
$y(0.5) = 1.86928,$ $y'(0.5) = 2.43464,$ $y''(0.5) = 4.21732$
$y(0.6) = 2.13383$

E5) Euler's method is $y_{k+1} = y_k + hf_k = y_k + h(1-2x_k y_k)$
$y(0.4) = 0.1948 + (0.2)(1-2 \times 0.2 \times 0.1948)$
$= 0.379216.$

E6) $y' = \frac{1}{x^2 + y},$ $y(4) = 4,$ $y'(4) = 0.05$
$y(4.1) = y(4) + hy'(4)$
$= 4 + (0.1)(0.05) = 4.005.$

E7) Euler's method $y' = (y-x)/(y+x)$, $y(0) = 1$, $y'(0) = 1$
$y(0.1) = 1 + (0.1)(1) = 1.1$

E8) Euler's method is
$y_{k+1} = h + y_k + hy_k^2$

Starting with $t_0 = 0$ and $y_0 = 1$, we have the following tables of values

Table 5: h = 0.2

T	Y(t)
0.2	1.4
0.4	1.992
0.6	2.986

$\therefore$ $y(0.6) = 2.9856$

Table 6: h = 0.1

T	Y(t)
0.1	1.2
0.2	1.444
0.3	1.7525
0.4	2.1596
0.5	2.7260
0.6	3.5691

$\therefore$ $y(0.6) = 3.5691$

E9) $y_{k+1} = y_k + h\ (t_k + y_k) = (1 + h)\ y_k + ht_k$

Starting with $t_0 = 0$, $y_0 = 1$, we obtain the following of values.

Table 7: h = 0.2

T	Y(t)
0.2	1.2
0.4	1.48
0.6	1.856

$\therefore$ y(0.6) = 1.856 with h = 0.2

Table 8: h = 0.1

t	y(t)
0.1	1.1
0.2	1.22
0.3	1.362
0.4	1.5282
0.5	1.72102
0.6	1.943122

$\therefore$ y(0.6) = 1.943122 with h = 0.1

Table 9: h = 0.05

t	y(t)	T	y(t)
0.05	1.05	0.35	1.46420
0.10	1.105	0.40	1.55491
0.15	1.16525	0.45	1.65266
0.20	1.23101	0.50	1.75779
0.25	1.30256	0.55	1.87068
0.30	1.38019	0.60	1.99171

$\therefore$ y(0.6) = 1.99171 with h = 0.05.

UNIT 4 SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS USING RUNGE-KUTTA METHODS

Structure **Page Nos.**

4.0 INTRODUCTION

In unit 3, we considered the IVPs

$$y' = f(t, y), \qquad y'(t_0) = y_0 \qquad (1)$$

and developed Taylor series method and Euler's method for its solution. As mentioned earlier, Euler's method being a first order method, requires a very small step size for reasonable accuracy and therefore, may require lot of computations. Higher order Taylor series require evaluation of higher order derivates either manually or computationally. For complicated functions, finding second, third and higher order total derivatives is very tedious. Hence, Taylor series methods of higher order are not of much practical use in finding the solutions of IVPs of the form given by Eqn. (1).

In order to avoid this difficulty, at the end of nineteenth century, the German mathematician, Runge observed that the expression for the increment function Ø (t, y, h) in the single step methods [see Eqn. (24) of Sec. 7.3, Unit 7]

$$y_{n+1} = y_n + h\, ø\, (t_n, y_n, h) \qquad (2)$$

can be modified to avoid evaluation of higher order derivatives. This idea was further developed by Runge and Kutta (another German mathematician) and the methods given by them are known as Runge-Kutta methods. Using their ideas, we can construct higher order methods using only the function f(t, y) at selected points on each subinterval. We shall, in the next section, derive some of these methods.

4.1 OBJECTIVES

After going through this unit, you should be able to:

- State the basic idea used in the development of Runge-Kutta methods;
- Obtain the solution of IVPs using Runge-Kutta methods of second, third and fourth order, and
- Compare the solutions obtained by using Runge-Kutta and Taylor series methods.

4.2 RUNGE-KUTTA METHODS

We shall first try to discuss the basic idea of how the Runge-Kutta methods are developed.
Consider the $O(h^2)$ singlestep method

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2} y''_n \qquad (3)$$

If we write Eqn. (3) in the form of Eqn. (2) i.e., in terms of Ø $[t_n, y_n, h]$ involving partial derivatives of f(t, y), we obtain

$$Ø(t, y, h) = f(t_n, y_n) + \frac{h}{2}\left[f_t(t_n, y_n) + f(t_n, y_n)\, f_y(t_n, y_n)\right] \qquad (4)$$

Runge observed that the r.h.s. of Eqn. (4) can also be obtained using the Taylor series expansion of $f(t_n + ph, y_n + qhf_n)$ as

$$f(t_n + ph, y_n + qhf_n) \approx f_n + ph\, f_t(t_n, y_n) + qhf_n\, f_y(t_n, y_n), \qquad (5)$$

Taylor's expansion in two variables t, y

where $f_n = f(t_n, y_n)$

Comparing Eqns. (4) and (5) we find that $p = q = ½$ and the Taylor series method of $O(h^2)$ given by Eqn. (3) can also be written as

$$y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f_n\right) \qquad (6)$$

Since Eqn. (5) is of $O(h^2)$, the value of y_{n+1} in (6) has the TE of $O(h^3)$. Hence the method (6) is of $O(h^2)$ which is same as that of (3).

The advantage of using (6) over Taylor series method (3) is that we need to evaluate the function f(t, y) only at two points (t_n, y_n) and $\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f_n\right)$. We observe that $f(t_n, y_n)$ denotes the slope of the solution curve to the IVP (1) at (t_n, y_n). Further,

$f\left[t_n + \frac{h}{2}, y_n + \left(\frac{h}{2}\right) f_n\right]$ denotes an approximation to the slope of the solution curve at

the point $\left[t_n + \frac{h}{2}, y\left(t_n + \frac{h}{2}\right)\right]$ Eqn. (6) denotes geometrically, that the slope of the

solution curve in the interval $[t_n, t_{n+1}]$ is being approximated by an approximation to

the slop at the middle points $t_n + \frac{h}{2}$. This idea can be generalized and the slope of the solution curve in $[t_n, t_{n+1}]$ can be replaced by a weighted sum of slopes at a number of

points in $[t_n, t_{n+1}]$ (called off-step points). This idea is the basis of the Runge-Kutta methods.

Let us consider for example, the weighted sum of the slopes at the two points $[t_n, y_n]$ and $[t_n + ph, y_n + qhf_n]$, $0 < p, q < 1$ as

$$\phi(t_n, y_n, h) = W_1 f(t_n, y_n) + W_2 f[t_n + ph, y_n + qhf_n] \qquad (7)$$

We call W_1 and W_2 as weights and p and q as scale factors. We have to determine the four unknowns W_1, W_2, p and q such that Ø (t_n, y_n, h) is of $O(h^2)$. Substituting Eqn. (5) in (7), we have

$$\phi(t_n, y_n, h) = W_1 f_n + W_2 [f_n + phf_t(t_n, y_n) + phf_n f_y(t_n, y_n)]. \qquad (8)$$

putting this in (2) we get,

$$\begin{aligned} y_{n+1} &= y_n + h\left[W_1 f_n + W_2\{f_n + phf_t(t_n, y_n) + qhf_n f_y(t_n, y_n)\}\right] \\ &= y_n + h(W_1 + W_2) f_n + h^2 W_2 (pf_t + qf_n f_y)_n \end{aligned} \qquad (9)$$

where $(\ )_n$ denotes that the quantities inside the brackets are evaluated at (t_n, y_n). Comparing the r.h.s. of Eqn. (9) with Eqn. (3), we find that

$$\left.\begin{aligned} W_1 + W_2 &= 1 \\ W_2 p = W_2 q &= \frac{1}{2} \end{aligned}\right\} \qquad (10)$$

In the system of Eqns. (10), since the number of unknowns is more than the number of equations, the solutions is not unique and we have infinite number of solutions. The solution of Eqn. (10) can be written as

$$\begin{aligned} W_1 &= 1 - W_2 \\ p = q &= 1/(2W_2) \end{aligned} \qquad (11)$$

By choosing W_2 arbitrarily we may obtain infinite number of second order Runge-Kutta methods. If $W_2 = 1$, $p = q = \frac{1}{2}$ and $W_1 = 0$, then we get the method (6). Another choice is $W_2 = \frac{1}{2}$ which gives $p = q = 1$ and $W_1 = \frac{1}{2}$. With this choice we obtain from (7), the method

$$y_{n+1} = y_n + \frac{h}{2}\left[f(t_n, y_n) + f(t_n + h, y_n + hf_n)\right] \qquad (12)$$

which is known as **Heun's method.**

Note that when f is a function of t only, the method (12) is equivalent to the trapezoidal rule of integration, whereas the method (6) is equivalent to the midpoint rule of integration. Both the methods (6) and (12) are of $O(h^2)$. The methods (6) and (12) can easily be implemented to solve the IVP (1). **Method** (6) is usually known as **improved tangent method** or **modified Euler method**. Method (12) is also known as **Euler – Cauchy method**.

We shall now discuss the Runge-Kutta methods of $O(h^2)$, $O(h^3)$ and $O(h^4)$.

4.2.1 Runge-Kutta Methods of Second Order

The general idea of the Runge-Kutta (R-K) methods is to write the required methods as

$y_{n+1} = y_n + h$ (weighted sum of the slopes).

$$= y_n + \sum_{i=1}^{m} W_i K_i \qquad (13)$$

where m slopes are being used. These slopes are defined by
$K_1 = hf(t_n, y_n)$,
$K_2 = hf(t_n + C_2h, y_n + a_{21}K_1)$,
$K_3 = hf(t_n + C_3h, y_n + a_{31}K_1 + a_{32}K_3)$,
$K_4 = hf(t_n + C_4h, y_n + a_{41}K_1 + a_{42}K_2 + a_{43}K_3)$,
etc. In general, we can write

$$K_i = hf\left[t_n + C_i h, \sum_{j=1}^{i-1} a_{ij}K_j\right], \; i = 1, 2, \ldots, m \text{ with } C_1 = 0 \qquad (14)$$

The parameters C_i, a_{ij}, W_j are unknowns and are to be determined to obtain the Runge-Kutta methods.

We shall now derive the second order Runge-Kutta methods.

Consider the method as

$$Y_{n+1} = y_n + W_1K_1 + W_2K_2, \qquad (15)$$

where

$$K_1 = hf(t_n, y_n)$$
$$K_2 = hf(t_n + C_2h, y_n + a_{21}K_1), \qquad (16)$$

where the parameters C_2, a_{21}, $W_1$and W_2 are chosen to make y_{n+1} closer to $y(t_{n=1})$.

The exact solution satisfies the Taylor series

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \frac{h^2}{2}y''(t_n) + \frac{h^3}{6}y'''(t_n) + \ldots. \qquad (17)$$

where

$$y' = f(t, y)$$
$$y'' = f_t + ff_y$$
$$y''' = f_{tt} + 2ff_{ty} + f_{yy}f^2 + f_y(f_t + ff_y)$$

We expand K_1 and K_2 about the point (t_n, y_n)

$$K_1 = hf(t_n, y_n) = hf_n$$

$K_2 = hf(t_n + C_2h, y_n + a_{21}hf_n)$

$$h\left\{f(t_n, y_n) + \left(C_2hf_t + a_{21}hf_nf_y\right) + \frac{1}{2!}\left(C_2^2h^2f_{tt} + 2C_2a_{21}h^2f_nf_{ty} + a_{21}^2h^2f_n^2f_{yy}\right) + \ldots\right\}$$

Substituting these values of K_1 and K_2 in Eqn. (15), we have

$$y_{n+1} = y_n + (W_1 + W_2)hf_n + h^2\left[W_2C_2f_t + W_2a_{21}f_nf_y\right]$$

$$+\frac{h^3}{2}W_2\left(C_2^2f_{tt} + 2C_2a_{21}f_nf_{ty} + a_{21}^2f_n^2f_{yy}\right) + \ldots. \qquad (18)$$

Comparing Eqn. (18) with (17), we have

$$W_1 + W_2 = 1$$
$$C_2W_2 = \frac{1}{2}$$
$$a_{21}W_2 = \frac{1}{2}$$

From these equations we find that if C_2 is chosen arbitrarily we have

$$a_{21} = C_2, W_2 = 1/(2C_2), \qquad W_1 = 1 - 1/(2C_2) \qquad (19)$$

The R-K method is given by

$$y_{n+1} = y_n + h\left[W_1f(t_n, y_n) + W_2f(t_n + C_2h, y_n + C_2hf_n)\right]$$

and Eqn. (18) becomes

$$y_{n+1} = y_n + hf_n\frac{h^2}{2}\left(f_1 + f_nf_y\right) + \frac{C_2h^3}{4}\left(f_{tt} + 2f_nt_{yy} + f_n{}^2f_{yy}\right) + \ldots.. \qquad (20)$$

Subtracting Eqn. (20) from the Taylor series (17), we get the truncation error as

$$TE = y(t_{n+1}) - y_{n+1}$$

$$h^3\left[\left(\frac{1}{6} - \frac{C_2}{4}\right)\left(f_{tt} + 2f_nf_{ty} + f_n^2f_{yy}\right) + \frac{1}{6}fy\left(f_t + f_nf_y\right)\right] + \ldots..$$

$$= \frac{h^3}{12}\left[(2 - 3C_2)y''' + 3C_2f_yy''\right] + \ldots.. \qquad (21)$$

Since the TE is of $O(h^3)$, all the above R-K methods are of second order. Observe that no choice of C_2 will make the leading term of TE zero for all f(t, y). The local TE depends not only on derivatives of the solution y(t) but also on the function f(t, y). This is typical of all the Runge-Kutta methods. Generally, C_2 is chosen between 0 and 1 so that we are evaluating f(t, y) at an off-step point in $[t_n, t_{n+1}]$. From the definition, every Runge-Kutta formula must reduce to a quadrature formula of the same order or greater if f(t, y) is independent of y, where W_i and C_i will be weights and abscissas of the corresponding numerical integration formula.

Best way of obtaining the value of the arbitrary parameter C_2 in our formula is to

i) choose some of W_i's zero so as to minimize the computations.
ii) choose the parameters to obtain least TE,
iii) choose the parameter to have longer stability interval.

Methods satisfying either of the condition (ii) or (iii) are called optimal **Runge-Kutta methods**.

We made the following choices:

i) $C_2 = \frac{1}{2}, \quad \therefore a_{21} = \frac{1}{2}, W_1 = 0, W_2 = 1,$ then

$$y_{n+1} = y_n + K_2,$$

$$K_1 = hf(t_n, y_n),$$

$$K_2 = hf\left(t_n + \frac{h}{2}, y_n + \frac{K_1}{2}\right) \qquad (22)$$

which is the same as **improved tangent** or **modified Euler's method.**

ii) $C_2 = 1, \quad \therefore a_{21} = 1, W_1 = W_2 = \frac{1}{2},$ then

$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2),$$

$$K_1 = hf(t_n, y_n),$$

$$K_2 = hf(t_n + h, y_n + K_1) \qquad (23)$$

which is same as the **Euler-Cauchy method.**

iii) $C_2 = \frac{2}{3}, \quad \therefore a_{21} = \frac{2}{3}, W_1 = \frac{1}{4}, W_2 = \frac{3}{4},$ then

$$y_{n+1} = y_n + \frac{1}{4}(K_1 + 3K_2),$$

$$K_1 = hf(t_n, y_n),$$

$$K_2 = hf\left(t_n + \frac{2h}{3}, y_n + \frac{2K_1}{3}\right) \qquad (24)$$

which is the **optimal R-K method.**

Method (24) is optimal in the sense that it has minimum TE. In other words, with the above choice of unknowns, the leading term in the TE given by (21) is minimum. Though several other choices are possible, we shall limit our discussion with the above three methods only.

In order to remember the weights W and scale factors C_i and a_{ij} we draw the following table:

C_2	a_{21}
	W_1 W_2

General form

1/2	1/2
	0 1

Improved tangent method

1	1
	1/2 1/2

Heun's method

2/3	2/3
	1/4 ¼

Optimal method

We now illustrate these methods through an example.

Example 1: Solve the IVP $y' = -ty^2, y(2) = 1$ and find y(2.1) and y(2.2) with h = 0.1 using the following R-K methods of $O(h^2)$

a) Improved tangent method [modified Euler method (22)]
b) Heun's method [Euler-Cauchy method (23)]
c) Optimal R-K method [method (24)]
d) Taylor series method of $O(h^2)$.

Compare the results with the exact solution

$$y(t) = \frac{2}{t^2 - 2}$$

Solution: We have the exact values
y(2.1) = 0.82988 and y(2.2) = 0.70422

a) Improved tangent method is

$$y_{n+1} = y_n + K_2$$

$$K_1 = hf(t_n, y_n)$$

$$K_2 = hf\left(t_n + \frac{h}{2}, y_n + \frac{K_1}{2}\right).$$

For this problem f(t, y) = – t y^2 and

$$K_1 = (0.1)[(-2)(1)] = -0.2$$

$$K_2 = (0.1)[(-2.05)(1 - 0.1)^2] = -0.16605$$

y(2.1) = 1-0.16605 = 0.83395

Taking $t_1 = 2.1$ and $y_1 = 0.83395$, we have

$$K_1 = hf(t_1, y_1) = (0.1)[(-2.1)(0.83395)^2] = -0.146049$$

$$K_2 = hf\left(t_1 + \frac{h}{2}, y_1 + \frac{k_1}{2}\right).$$

$$= (0.1)[-(2.15)(0.83395 - 0.0730245)^2] = -0.124487$$

$y(2.2) = y_1 + k_2 = 0.83395 - 0.124487 = 0.70946$

b) Heun's method is :

$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2)$$

$$K_1 = hf(t_n, y_n) = -0.2$$

$$K_2 = hf(t_n + h, y_n + K_1) = -0.1344$$

y(2.1) = 0.8328
Taking $t_1 = 2.1$ and $y_1 = 0.8328$, we have
$K_1 = -0.14564$, $K_2 = -0.10388$
y(2.2) = 0.70804

c) Optimal method is:

$$y_{n+1} = y_n + \frac{1}{4}(K_1 + 3K_2)$$

$$K_1 = hf(t_n, y_n) = -0.2$$
$$K_2 = hf\left(t_n + \frac{2h}{3}, y_n + \frac{2k_1}{3}\right) = 0.15523$$

y(2.1) = 0.83358

Taking $t_1 = 2.1$ and $y_1 = 0.83358$, we have
$K_1 = -0.1459197$, $K_2 = -0.117463$

y (2.2) = 07090

d) Taylor series method of $O(h^2)$:

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2}y''_n$$

$$y' = -ty^2, y'' = -y^2 - 2tyy'$$

y(2) = 1, y' (2) = -2, y'' (2) = 7
y(2.1) = 0.8350

with $t_1 = 2.1$, $y_1 = 0.835$, we get
y ′ (2.1) = -1.4641725, y ′′ (2.1) = 4.437627958
y(2.2) = 0.71077

We now summarise the results obtained and give them in Table 1.

Table 1

Solution and errors in solution of $y' = -ty^2$, y (2) = 1, h = 0.1. Number inside brackets denote the errors.

T	Method (22)	Method (23)	Method (24)	Method Taylor $O(h^2)$	Exact Solution
2.1	0.83395 (0.00405)	0.8328 (0.0029)	0.83358 (0.00368)	0.8350 (0.0051)	0.8299
2.2	0.70746 (0.0033)	0.708.4 (0.00384)	0.7090 (0.0048)	0.71077 (0.00657)	0.7042

You may observe her that all the above numerical solutions have almost the same error. You may now try the following exercises:

Solve the following IVPs using Heun's method of $O(h^2)$ and the optimal R-K method of $O(h^2)$.

E1) 10y' = t2+y2, y(0) = 1. find y (0.2) taking h = 0.1.
E2) y' = 1+y2, y(0) = 0. Find y(0.4) taking h = 0.2. iven that the exact solution is y(t) = tan t, find the errors.

Also compare the errors at t = 0.4, obtained here with the one obtained by Taylor series method of O(h2)

E3) $y' = 3t + \frac{1}{2}y, y(0) = 1$. Find y(0.2) taking h = 0.1. Given $y(t) = 13e^{t/2} - 6t - 12$, find the error.

Let us now discuss the R-K methods of third order.

4.2.2 Runge-Kutta Methods of Third Order

Here we consider the method as

$$y_{n+1} = y_n + W_1K_1 + W_2K_2 + W_3K_3 \qquad (25)$$

where

$K_1 = h\, f(t_n, y_n)$

$K_2 = hf\,(t_n + C_2h,\ y_n + a_{21}\, K_1)$

$K_3 = hf\,(t_n + C_3h,\ Y_n + a_{31}K_1 + a_{32}\, K_2)$

Expanding K_2, K_3 and Y_{n+1} by Taylor series, substituting their values in Eqn. (25) and comparing the coefficients of powers of h, h^2 and h^3, we obtain

$$a_{21} = C_2 \qquad C_2W_2 + C_3W_3 = \frac{1}{2}$$

$$a_{31} + a_{32} = C_3 \qquad C\frac{2}{2}W_2 + C\frac{2}{3}W_3 = \frac{1}{3} \qquad (26)$$

$$W_1 + W_2 + W_3 = 1 \qquad C_2a_{32}W_3 = \frac{1}{6}$$

We have 6 equations to determine the 8 unknowns (3 W′s + 2C′s + 3a′s). Hence the system has two arbitrary parameters. Eqns. (26) are typical of all the R-K methods. Looking at Eqn. (26), you may note that the sum of a_{ij}'s in any row equals the corresponding C_i's and the sum of the Wi's is equal to 1. Further, the equations are linear in w_2 and w_3 and have a solution for W_2 and W_3 if and only if

$$\begin{vmatrix} C_2 & C_3 & -1/2 \\ C^2 & C^2 & -1/3 \\ 0 & C_{2^{a32}} & -1/6 \end{vmatrix} = 0$$

(Ref. Sec. 8.4.2, Unit 8, Block-2, MTE-02, IGNOU material).

Expanding the determinant and simplifying we obtain

$$C_2(2 - 3C_2)\,a_{32} - C_3(C_3 - C_2) = 0, C_2 \neq 0 \qquad (27)$$

Thus we choose C_2, C_3 and a_{32} satisfying Eqns. (27).

Since two parameters of this system are arbitrary, we can choose C_2, C_3 and determine a_{32} from Eqn. (27) as

$$a_{32} = \frac{C_3(C_3 - C_2)}{C_2(2 - 3C_2)}$$

If $C_3 = 0$, or $C_2 = C_3$ then $C_2 = \frac{2}{3}$ and we can choose $a_{32} \neq 0,$ arbitrarily. All C_i's should be chosen such that $0 < C_i < 1$. Once C_2 and C_3 are prescribed, W_i's and a_{ij}'s can be determined from Eqns. (26).

We shall list a few methods in the following notation

C_2	a_{21}		
C_3	A_{31}	A_{32}	
	W_1	W_2	W_3

i) Classical third order R-K method

1/2	1/2		
1	-1	2	
	1/6	4/6	1/6

$$Y_{n+1} = Y_n + \tfrac{1}{6}\left(K_1 + 4k_2 + K_3\right) \qquad (28)$$

$$K_1 = hf(t_n, y_n)$$

$$K_2 = hf\left(t_n + \frac{h}{2}, y_n + \frac{K_1}{2}\right)$$

$$K_3 = hf(t_n - \frac{h}{2}; y_n - K_1 + 2K_2)$$

ii) Heun's Method

1/3	1/3		
2/3	0	2/3	
	1/4	0	3/4

$$y_{n+1} = y_n + \frac{1}{4}(K_1 + 3K_3) \qquad (29)$$

$$K_1 = h\, f(t_n, y_n)$$

$$K_2 = h\, f\left(t_n + \frac{h}{3}, y_n + \frac{K_1}{3}\right)$$

$$K_3 = h\, f\left(t_n + \frac{2h}{3}, y_n + \frac{2K_2}{3}\right)$$

iii) Optimal method

½	1/2		
¾	0	¾	
	2/9	3/9	4/9

$$y_{n+1} = y_n + \frac{1}{9}(2K_1 + 3K_2 + 4K_3) \qquad (30)$$

$$K_1 = h\, f(t_n, y_n),$$

$$K_2 = f(t_n + \frac{h}{2}, y_n + \frac{K_1}{2}),$$

$$K_3 = h\, f\left(t_n + \frac{3h}{4}, y_n + \frac{3K_2}{4}\right).$$

We now illustrate the third order R-K methods by solving the problem considered in Example 1, using (a) Heun's method (29) (b) optimal method (30).

a) Heun's method

$$y_{n+1} = y_n + \frac{1}{4}(K_1 + 3K_3)$$

$$K_1 = h\,f(t_n, y_n)$$

$$= -0.2$$

$$K_2 = h\,f\left(t_n + \frac{h}{3}, y_n + \frac{K_1}{3}\right)$$

$$= -0.17697$$

$$K_3 = h\,f\left(t_n + \frac{2h}{3}, y_n + \frac{2k_2}{3}\right)$$

$$= -0.16080$$

$$y(2.1) = 0.8294$$

Taking $t_1 = 2.1$ and $y_1 = 0.8294$, we have

$$K_1 = -0.14446$$

$$K_2 = -0.13017$$

$$K_3 = -0.11950$$

$$y(2.2) = 0.70366$$

b) Optimal method

$$y_{n+1} = y_n + \frac{1}{9}(2K_1 + 3K_2 + 4K_3)$$

$$K_1 = -0.2$$

$$K_2 = -0.16605$$

$$K_3 = -0.15905$$

$$y(2.1) = 0.8297$$

Taking $t_1 = 2.1$ and $y_1 = 0.8297$, we have

$$K_1 = -0.14456$$

$$K_2 = -012335$$

$$K_3 = -0.11820$$

$$y(2.2) = 0.70405$$

You can now easily find the errors in these solutions and compare the results with those obtained in Example 1.
And now here is an exercise for you.

E4) Solve the IVP

Y' = y-t y (0) =2

Using third order Heun's and optimal R-K methods. Find (02) taking h = 0.1. Given the exact solution to be y (t) = 1+t+e^t, find the errors at t = 0.2.

We now discuss the fourth order R-K methods.

4.2.3 Runge-Kutta Methods of Fourth Order

Consider the method as

$$\begin{aligned} y_{n+1} &= y_n + W_1K. + W_2K_2 + W_3K_3 + W_4K_4 \\ K_1 &= h\,f(t_n, y_n), \\ K_2 &= h\,f(t_n + C_2h, y_n + a_{21}K_1), \\ K_3 &= h\,f(t_n + C_3h, y_n + a_{31}K_1 + a_{32}K_2), \\ K_4 &= h\,f(t_n + C_4h, y_n + a_{41}K_2 + a_{43}K_3). \end{aligned} \qquad (31)$$

Since the expansions of K_2, K_3, K_4 and y_{n+1} in Taylor series are completed, we shall not write down the resulting system of equations for the determination of the unknowns. It may be noted that the system of equations has 3 arbitrary parameters. We shall state directly a few R-K methods of $O(h^4)$. The R-K methods (31) can be denoted by

C_2	a_{21}			
C_3	a_{31}	a_{32}		
C_4	a_{41}	a_{42}	a_{43}	
	W_1	W_2	W_3	W_4

For different choices of these unknowns we have the following methods :

i) **Classical R-K method**

½	1/2			
½	0	½		
1	0	0	1	
	1/6	2/6	2/6	1/6

$$
\begin{aligned}
&y_{n+1} = y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4),\\
&K_1 = h\,f(t_n, y_n),\\
&K_2 = h\,f\left(t_n \frac{h}{2}, y_n + \frac{K_1}{2}\right),\\
&K_3 = h\,f\left(t_n + \frac{h}{2}, y_n + \frac{K_2}{2}\right),\\
&K_4 = h\,f(t_n + h, y_n + K_3).
\end{aligned}
\qquad (32)
$$

This is the widely used method due to its simplicity and moderate order. We shall also be working out problems mostly by the classical R-K method unless specified otherwise.

1/2	1/2			
1/2	$\frac{(\sqrt{2}-1}{2}$	$\frac{(2-\sqrt{2}}{2}$		
1	0	$\frac{\sqrt{2}}{2}$	$1+\frac{\sqrt{2}}{2}$	
	1/6	$\frac{(2-\sqrt{2}}{6}$	$\frac{(2+\sqrt{2}}{6}$	1/6

$$
\begin{aligned}
&y_{n+1} = y_n + \frac{1}{6}(K_1 + (2-\sqrt{2})\,K_2 + (2+\sqrt{2})\,K_3 + K_4) \qquad (33)\\
&K_1 = h\,f(t_n, y_n),\\
&K_2 = h\,f\left(t_n + \frac{h}{2}, y_n + \frac{K_1}{2}\right),\\
&K_3 = h\,f\left(t_n + \frac{h}{2}, y_n + \left(\frac{\sqrt{2}-1}{2}\right)K_1 + \left(\frac{2-\sqrt{2}}{2}\right)K_2\right),\\
&K_4 = h\,f\left(t_n + h, y_n - \frac{\sqrt{2}}{2}K_2 + \left(1+\frac{\sqrt{2}}{2}\right)K_3\right),
\end{aligned}
$$

The Runge-Kutta-Gill method is also used widely. But in this unit, we shall mostly work out problems with the classical R-K method of $O(h^4)$. Hence, whenever we refer to R-K method of $O(h^4)$ we mean only the classical R-K method of $O(h^4)$ given by (32). We shall now illustrate this method through examples.

Example 2 : Solve the IVP $y' = t+y$, $y(0) = 1$ by Runge-Kutta method of $O(h^4)$ for $t \in (0,0.5)$ with $h = 0.1$. Also find the error at $t = 0.5$, if the exact solution is $y(t) = 2e^t - t - 1$.

Solution : We use the R-K method of $O(h^4)$ given by (32).

Initially, $t_0 =$, $t_0 = 1$

 We have

$$K_1 = hf(t_0, y_0) = (0.1)[0+1] = 0.1$$

$$K_2 = hf\left(t_0 + \frac{h}{2}, y_0 + \frac{K_1}{2}\right) = (0.1)[0.05 + 1 + 0.05] = 0.11$$

$$K_3 = hf\left(t_0 + \frac{h}{2}, y_0 + \frac{K_2}{2}\right) = (0.1)[0.05 + 1 + 0.055] = 0.1105$$

$$K_4 = hf(t_0 + h, y_0 + K_3) = (0.1)[0.1 + 1 + 0.1105] = 0.12105$$

$$y_1 = y_0 + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4)$$

$$= 1 + \frac{1}{6}\,[1 + 0.22 + 0.2210 + 0.12105] = 1.11034167$$

Taking $t_1 = 0.1$ and $y_1 = 1.11034167$, we repeat the process.

$$K_1 = hf(t_1, y_1) = (0.1)\,[0.1 + 1.11034167] = 0.121034167$$

$$K_2 = hf\left(t_1 + \frac{h}{2}, y_1 + \frac{K_1}{2}\right) = (0.1)$$

$$\left[0.1 + 0.05 + 1.1103416 + \frac{(0.121034167)}{2}\right] = 0.132085875$$

$$K_3 = hf\left(t_1 + \frac{h}{2}, y_1 + \frac{K_2}{2}\right) = (0.1)$$

$$\left[0.1 + 0.05 + 1.1103417 + \frac{(0.132085875)}{2}\right] = 0.132638461$$

$$K_4 = hf\,(t1 + h, y1 + K3) = (0.1)\,[0.1 + 0.05 + 1.11034167 + \frac{(0.132085875)}{2}$$

$$= 0.144303013$$

$$y_2 = y_1 + \frac{1}{6}\,(K_1 + 2K_2 + 2K_3 + K_4)$$

$$= 1.11034167 + \frac{1}{6}\,[(0.121034167 + 2(0.132085875) + 2(0.132638461)$$

$$+ 0.1443030131 = 1.24280514$$

Rest of the values y_3, y_4, y_5 we give in Table 2.

Table 2

t_n	y_n
0.0	1
0.1	1.11034167
0.2	1.24280514
0.3	1.39971699
0.4	1.58364848
0.5	1.79744128

Now the exact solution is

$$y(t) = 2e^t - t - 1$$

Error at t = 0.5 is

$$y(0.5) - y_5 = (2e^{0.5} - 0.5 - 1) - 1.79/44128$$
$$= 1.79744254 - 1.79744128$$
$$= 0.000001261$$
$$= 0.13 \times 10^{-5}.$$

Let us consider another example

Example 3 : Solve the IVP

$$y' = 2y + 3e^t, y(0) = 0 \text{ using}$$
a) classical $R - K$ method of $O(h^4)$
b) $R - K$ Gill method of $O(h^4)$,

Find y (0.1), y(0.2), y(0.3) taking h = 0.1. Also find the errors at t = 0.3, if the exact solution is $y(t) = 3(e^{2t} - e^t)$.

Solution: a) Classical R-K method is

$$y_{n+1} = y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4)$$

Here $t_0 = 0$, $y_0 = 0$, h = 0.1

$K_1 = h\, f(t_0, y_0) = 0.3$

$K_2 = h\, f\left(t_0 + \frac{h}{2}, y_0 + \frac{K_1}{2}\right) = 0.3453813289$

$K_3 = h\, f\left(t_0 + \frac{h}{2}, y_0 + \frac{K_2}{2}\right) = 0.3499194618$

$K_4 = h\, f\ (t_0 + h, y_0 + K_3) = 0.4015351678$

$y_1 = 0.3486894582$

Taking $t_1 = 0.1, y_1 = 0.3486894582,$ we repeat the process and obtain

$K_1 = 0.4012891671$	$K_2 = 0.4584170812$
$K_3 = 0.4641298726$	$K_4 = 0.6887058455$
$Y(0.2) = 0.8112570941$	

Taking $t_2 = 0.2$, $y_2 = 0.837870944$ and repeating the process we get

$K_1 = 0.53399502$	$K_2 = 0.579481565$
$K_3 = 0.61072997$	$K_4 = 0.694677825$
$\therefore$ y (0.3) = 1.416807999	

b) R-K gill method is

$$y_{n+1} = y_n + \frac{1}{6}\left(K_1 + (2 + (2 - \sqrt{2})K_2 + (2 + \sqrt{2})K_3 + K_4\right)$$

Taking $t_0 = 0$, $y_0 = 1$ and h = 0.1, we obtain

$K_1 = 0.3$	$K_2 = 0.3453813289$
$K_3 = 0.3480397056$	$K_4 = 0.4015351678$
y (0.1) = 0.3486894582	

Taking $t_1 = 0.1$, $y_1 = 0.3486894582$, we obtain

K_1= 0.4012891671	K_2 = 0.4584170812
K_3 = 0.4617635569	K_4 = 0.5289846936
y (0.2) = 0.8112507529	

Taking $t_2 = 0.2$, $y_2 = 0.8112507529$, we obtain

K_1= 0.528670978	K_2 = 0.6003248734
K_3 = 0.6045222614	K_4 = 0.6887058455
y (0.3) = 1.416751936	

From the exact solution we get

$y(0.3) = 1.416779978$

Error in classical R-K method (at t = 0.3) = 02802×10^{-04}

Error in R-K Gill method (at t = 0.3) = 0.2804×10^{-04}

You may now try the following exercises

Solve the following IVPs using R-K method of $O(h^4)$

E5) $y' = \frac{y - t}{y + t}, y(0) = 1$. Find y(0.5) taking h = 0.5.

E6) $y' \; 1 - 2ty$, $y(0.2) = 0.1948$. Find y(0.4) taking h = 0.2.

E7) $10ty' + y^2 = 0, y(4) = 1$. Find y(4.2) taking h = 0.2. Find the error give n the exact solution is $y(t) = \frac{1}{c + 0.1 \text{in } t'}$ where c = 0.86137

E8) $y' = \frac{1}{t^2} - \frac{y}{t} - y^2, y(1) = -1$. Find y (1.3) taking h = 0.1. Given the exact solution to be $y(t) = \frac{1}{t'}$ find the error at t = 1.3.

We now end this unit by giving a summary of what we have covered in it.

4.3 SUMMARY

In this unit we have learnt the following :

1) Runge-Kutta methods being singlestep methods are self-starting methods.

2) Unlike Taylor series methods, R-K methods do not need calculation of higher order derivatives of f(t, y) but need only the evaluation of f(t, y) at the off-step points.

3) For given IVP of the form

$y' = f(t, y), y'(t_0) = y_0, t \varepsilon [t_0, b]$

where the mesh points are $t_j = t_0 + jh, j = 0, 1, ..., n.$

$t_n = b = t_0 + nh, R - K$ methods are obtained by writing

$y_{n+1} = y_n + h$ (weighted sum of the slopes)

$$= y_n + \sum_{i=1}^{m} W_i K_i$$

where in slopes are used. These slopes are defined by

$$K_i = f\left[t_n + C_i h, \sum_{i=1}^{i-1} a_{ij} k_j\right], i = 1, 2, ..., m, C_1 = 0.$$

Here is the order of the method. The unknowns C_i, a_{ij} and W_j are then obtained by expanding K_i's and y_{n+1} in Taylor series about the point (t_n, y_n) and comparing the coefficients of different powers of h.

4.4 SOLUTIONS/ANSWERS

E1) Heun's method : $y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2)$

Starting with $t_0 = 0, y_0 = 1, h = 0.1$

$\therefore K_1 = 0.01$

$K_2 = 0.010301$

$y(0.1) = 1.0101505$

Taking $t_1 = 0.1$, $y_1 = 1.0101505$

$K_1 = 0.0103040403$
$K_2 = 0.0181327468$
$y(0.2) = 1.020709158$

Optimal R-K, method : $y_{n+1} = y_n + \frac{1}{4}(K_1 + 3K_2)$

$t_0 = 0, y_0 = 1, h = 0.1$
$K_1 = 0.01, K_2 = 0.01017823$
$y(0.1) = 1.010133673$
$t_1 = 0.1, y_1 = 1.010133673$
$K_i = 0.0103037, K_2 = 0.010620.$
$y(0.2) = 1.020675142$

E2) Heun's method:

$K_1 = 0.2$, $K_2 = 0.208$
$y(0.2) = 0.204$
$K_1 = 0.2083232$, $K_2 = 0.2340020843$
$y(0.4) = 0.4251626422$

Optimal R-K, method:
$K_1 = 0.2$, $K_2 = 0.2035556$
$y(0.2) = 0.2026667$
$K_1 = 0.2082148$, $K_2 = 0.223321245$
$y(0.4) = 0.42221134$

Taylor Series method
$y' = 1 + y^2$, $y'' = 2yy'$
$y(0) = 0$, $y'(0) = 1$, $y''(0) = 0$
$y(0.2) = 0.2$
$y'(0.2) = 1.04$, $y''(0.2) = 0.416$
$y(0.4) = 0.41632$

74Now the exact solution is $y(t) = \tan t$

Exact y (0.4) = 0.422793219

Error in Heun's method = 0.422793219

Error in Optimal R-K method = 0.236×10^{-2}

Error in Optimal R-K method = 0.582×10^{-3}

Error in Taylor series method = 0.647×10^{-2}

E3) Heun's method:

$K_1 = 0.05, \quad K_2 = 0.0825$
$y(0.1) = 1.06625$
$K_1 = 0.0833125$
$y(0.2) = 1.166645313$

Optimal R-K, method

$K_1 = 005, \quad K_2 = 0.071666667$
$y(0.2) = 1.166645313$
Exact $y(0.2) = 1.167221935$

Error in both the methods is same and = 0.577×10^{-3}

E4) Heun's method : $y_{n+1} = y_n + \frac{1}{4}(K_1 + 3K_3)$

Starting with $t_0 = 0$, $y_0 = 2$, h = 0.1, we have

$K_1 = 0.2, \quad K_2 = 0.203334, K_3 = 0.206889$
$y(0.1) = 2.205167$
$t_1 = 0.1, \; y_1 = 2.205167$ we have
$K_1 = 0.210517, \quad K_2 = 0.214201, \; K_3 = 0.218130$
$y(0.2) = 2.421393717$

Optimal $R-K$ method: $y_{n+1} = y_n + \frac{1}{9}(2K_1 + 3K_2 + 4K_3)$

$K_1 = 0.2, \quad K_2 = 0.205, \; K_3 = 0.207875$
$y(0.1) = 2.205167$
$t_1 = 0.1, y_1 = 2.205167$
$K_1 = 0.2105167, \quad K_2 = 0.2160425, \; K_3 = 0.219220$
$y(0.2) = 2.421393717$
exact $y(0.2) = 2.421402758$
Since $y(0.2)$ is same by both the mehods

Error $= 0.9041 \times 10^{-5}$ in both the methods at t = 0.2.

$K_1 = 0.5, \quad K_2 = 0.333333$

E5) $K_3 = 0.3235294118, \quad K_4 = 0.2258064516$
$y(0.5) = 1.33992199.$

E6) $K_1 = 0.184416, \quad K_2 = 0.16555904$

$K_3 = 0.1666904576$, $K_4 = 0.1421615268$
y (0.4) = 0.3599794203.

E7) $K_1 = -0.005$, $K2 = -0.004853689024$
$K_3 = -0.0048544$, $K4 = -0.004715784587$
y (4.2) = 0.9951446726.
Exact y (4.2) = 0.995145231, Error = 0.559×10^{-6}

E8) K_1 0.1, $K_2 = 0.09092913832$
$K_3 = 0.9049729525$, $K_4 = 0.8260717517$
y(1.1) = – 0.909089993
$K_1 = 0.08264471138$, $K_2 = 0.07577035491$
$K_3 = 0.07547152415$, $K_4 = 0.06942067502$
y(1.2) = – 0.8333318022
$K_1 = 0.06944457204$, $K_2 = 0.06411104536$
$K_3 = 0.06389773475$, $K_4 = 0.0591559551$
y(1.3) = – 0.7692307692
Exact y(1.3) = – 0.7692287876
Error = 0.19816 x 10^{-5}

UNIT 1 PROBABILITY DISTRIBUTIONS

Structure **Page Nos.**

1.0 INTRODUCTION

The discipline of statistics deals with the collection, analysis and interpretation of data. Outcomes may vary even when measurements are taken under conditions that appear to be the same. Variation is a fact of life. Proper statistical methods can help us understand the inherent variability in the collected data, and facilitate the appropriate analysis of the same.

Because of this variability, uncertainty is a part of many of our decisions. In medical research, for example, interest may center on the effectiveness of a new vaccine for AIDS; an agronomist may want to decide if an increase in yield can be attributed to a new strain of wheat; a meteorologist may be interested in predicting whether it is going to rain on a particular day; an environmentalist may be interested in testing whether new controls can lead to a decrease in the pollution level; an economist's interest may lie in estimating the unemployment rate, etc. Statistics, and probabilistic foundations on which statistical methods are based, can provide the models that may be used to make decisions in these and many other situations involving uncertainties.

Any realistic model of a real world phenomenon must take into account the possibilities of randomness. That is, more often that not, the quantities we are interested in will not be predicted in advance, but rather will exhibit as inherent variation that should be taken into account by the model. Such a model is, naturally enough, referred to as a probability model.

In this unit we shall see what is a random variable and how it is defined for a particular random experiment. We shall see that there are two major types of probability distribution. We shall investigate their properties and study the different applications.

1.1 OBJECTIVES

After reading this unit, you should be able to

- describe events and sample spaces associated with an experiment;
- define a random variable associated with an experiment;
- decide whether a random variable is discrete or continuous;
- describe the following distributions
  a) Binomial distribution
  b) Poisson distribution
  c) Uniform distribution
  d) Exponential distribution
  e) Normal distribution
  f) Chi-square distribution.

1.2 RANDOM VARIABLES

Definition A "Random experiment" or a "Statistical experiment" is any act whose outcome can not be predicted in advance. Any outcome of a random experiment is known as "event"

We will start with the following illustrations :

1) The number of telephone calls received by Monica, a telephone operator in a call center in Delhi, between say 1:00 am and 3:00 am in the early morning.
2) The amount of rainfall in Mumbai on August 1st.
3) The number of misprints on a randomly chosen page of a particular book.
4) The final results of the 5 one-day matches between India-Pakistan.
5) The outcome of rolling die.
6) The volume of sales of a certain item in a given year.
7) Time to failure of a machine.

In all the above cases there is one common feature. These experiments describe the process of associating a number to an outcome of the experiment (i.e. to an event). A function which associates a number to each possible outcome of the experiment is called a "random variable". It is often the case that our primary interest is in the numerical value of the random variable rather than the outcome itself. The following examples will help to make this idea clear.

EXAMPLE 1: Suppose we are interested in the number of heads, say X, obtained in three tosses of a coin.
SOLUTION
If we toss a coin three times, then the experiment has a total of eight possible outcomes, and they are as follows;

a_1={HHH}, a_2={HHT}, a_3={HTH}, a_4={HTT}
a_5={THH}, a_6={THT}, a_7={TTH}, a_8={TTT}

Denoting the event corresponding to getting k heads, k=0,1,2,3,.. as{X=k}, observe that

{X=0} => {a_8} ; {X=1} => {a_4, a_6,a_7} ; {X=2} =>{a_2, a_3,a_5} ; {X=1} =>{a_1}

Note that each value in the support of X corresponds to some element (or set of elements) in the sample space **S.** For example the , the value 0 corresponds to the element $\{a_8\}$, while the value 1 corresponds to the set of elements $\{ a_4, a_6, a_7\}$

Therefore the sample space **S,** the set of all possible outcomes of this experiment can be expressed as

$$\mathbf{S}=\{a_1, a_2, \ldots, a_8\}$$

Since X is the characteristic, which denotes the number of heads out of the three tosses, it is associated with each outcome of this experiment. Therefore, X is a function defined on the elements of **S** and the possible values of *X* are $\{0,1,2,3\}$. The set of possible values that X can take is called the support of *X* which may be denoted as χ . Observe, that *X* can be explicitly expressed as follows;

$X(a_1)=3, X(a_2)=X(a_3)=X(a_5)=2, X(a_4)=X(a_6)=X(a_7)=1, X(a_8)=0$

It is interesting to observe that to each value there is always some element in sample space or a set of element in sample spaces. For, example, the set of element in sample spaces corresponding to the value '0' is the point $\{a_8\}$; for 1, the set is $\{a_4, a_6, a_7\}$, for 2, the set is $\{a_2, a_3, a_5\}$ and for 3 the point is $\{a_1\}$.

Therefore, we can easily make the following identification of events corresponding to the values associated by X. Denoting the event corresponding to '0', as {X=0}, similarly for other values, observe that

$\{X=0\}=\{a_8\}; \{X=1\}=\{a_4, a_6, a_7\}; \{X=2\}=\{a_2, a_3, a_5\}\ \{X=3\}=\{a_1\}$

If we assume that the coin is unbiased and the tosses have been performed independently, the probabilities of all the outcomes of the sample space are equal, that is $P(a_1)=P(a_2)=\ldots=P(a_8)=\frac{1}{8}$. Therefore, using the probability law of disjoint events we can easily obtain

$$P(\{X=0\}) = P(\{a_8\}) = \frac{1}{8}$$

$$P(\{X=1\}) = P(\{a_4, a_6, a_7\}) = P(\{a_4\})+P(\{a_6\})+P(\{a_7\}) = \frac{3}{8}$$

$$P(\{X=2\}) = P(\{a_2, a3, a5\}) = P(\{a_2\})+P(\{a_3\})+P(\{a_5\}) = \frac{3}{8}$$

$$P(\{X=3\}) = P(\{a_1\}) = \frac{1}{8}$$

Therefore, in this case the random variable X takes four values 0,1, 2,3 with the probabilities 1/8,3/8,3/8,1/8 respectively. It is also important to observe that

$$P(\{X=0\})+P(\{X=1\})+P(\{X=2\})+P(\{X=3\})=1$$

It is not a coincidence, it is always true. If we add the probabilities of all possible values of a random variable it is always one.

To sum up, we say that the random variable *X*, is a real valued function defined on all the elements of a sample space of a random experiment. The random variable takes different values, and for each value there is a probability associated with it. The sum of all the probabilities of all the points in the support χ of the random variable *X* adds up to one. The following figure will demonstrate the random variable *X*.

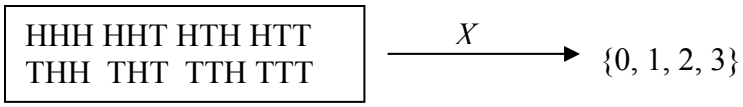


Figure 1: Schematic representation of a random variable

Because of this representation , one can define probabilities for the set of numbers(depending on the random variable) rather than working with arbitrary space and this simplifies the problem considerably. Now we consider the following example.

EXAMPLE 2: You have purchased a new battery operated wristwatch and you have inserted one new battery into it. Suppose you are interested in the following:

a) How long will it be before the first battery needs replacement ?
b) How many battery will have to be replaced during a one year period?

SOLUTION

Note that both (a) and (b) are random variables and let us discuss them one by one. In case (a), we want to find the duration of time before the battery needs to be replaced. Note that the variable takes values continuously along a line say from the time duration A to time duration B. No values in between A and B are left out. In other words there is no break in the values assumed by this random variable.

In case (b), the random variable is the number of batteries. This variable can take values 0 or 1 or 2 etc. There is no continuity, since only non-negative integer values can be assumed. So, the range of this variable is a discrete set of points. From this discussion it is clear that the random variable X defined in Example 1 is also a discrete random variable. The above examples shows that the random variables can be of two types. We will distinguish between the following two types of random variables;

1. Discrete Random Variable , and
2. Continuous Random Variable.

Check Your Progress 1

E 1: Suppose you take a 50-question multiple-choice examination, guessing your answer, and are interested in the number of correct answers obtained. Then

(a) What is the random variable X that you will consider for this situation?
(b) What is the set of possible values of X in this example?
(c) What does P(X=10) mean in this context?

Now in the next two sections we will describe the discrete and continuous random variables in detail.

1.3 DISCRETE RANDOM VARIABLE

In this section, we define properly a discrete random variable and mention some of its basic properties.

DEFINITION : A random variable X is said to be discrete, if the total number of values X can take is finite or countably infinite(i.e the support of X is either finite or countable).

The support χ of X may be listed as $\{a_0, a_1, a_2, \ldots\}$. Moreover, for each value of a_i, there is a probability associated with it. Suppose we denote them as $\{p_0, p_1, p_2, \ldots\}$,

therefore, we have P(X= a_i)=p_i for I=0,1,…. From the properties of a random variable and from the probability law, we have

(a) pi ≥ 0 for all i ≥ 0

(b) $\sum_{i=0}^{\infty} p_i = p_0 + p_1 + p_2 + ... = 1$

From the above discussions, it follows that there exists a function p: $\chi \rightarrow \mathbf{R}$ as follows;

$$p(a) = \begin{cases} p_i & if \quad a = a_i; \quad i = 0,1,2..... \\ 0 & \text{otherwise} \end{cases}$$

This function p is called the probability mass function (p.m.f.) of the discrete random variable X.
The collection of the pairs $\{(a_i, p_i; I=0,1,\ldots\}$ is called the probability distribution of X.

Another function which plays a very important role for any random variable is known as the cumulative distribution function (c.d.f.) or simply the distribution function of the random variable. The c.d.f. F : $\mathbf{R} \rightarrow [0,1]$ of the random variable X is defined as

$$F(b) = P(X \leq b), \text{ for } -\infty < b < \infty.$$

In other words, F(b) denotes the probability that the random variable X takes on a value which will be less than or equal to b. Some important properties of the c.d.f. F (.) are
(a) F (b) is a non-decreasing function of b.
(b) $\lim_{b \to \infty} F(b) = 1$
(c) $\lim_{b \to -\infty} F(b) = 0$

Now we clarify these concepts with the same example discussed in the previous section. Suppose X is the random variable denoting the number of heads obtained in three independent tosses of a fair coin, then the probability mass function (p.m.f) p is the function, p: $\chi \rightarrow \mathbf{R}$, such that

$$p(0) = \frac{1}{8}, \; p(1) = p(2) = \frac{3}{8}, \; p(3) = \frac{1}{8}$$

Therefore, $p(a_i) = p_i \geq 0$, for all a_i and

$$\sum_{i=0}^{3} p_i = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$$

In this case the p.m.f of the random variable by the function p and the corresponding probability distribution is the set $\left\{\left(0,\frac{1}{8}\right),\left(1,\frac{3}{8}\right),\left(2,\frac{3}{8}\right),\left(3,\frac{1}{8}\right)\right\}$ .This can also be expressed in a tabular form as follows

TABLE 1
PROBABILITY DISTRIBUTION OF THE NUMBER OF HEADS IN THREE INDEPENDENT TOSSES OF A FAIR COIN

The number of heads (X value)	Probability
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$

3	$\frac{1}{8}$

Now let us see the graphical representation of the distribution.

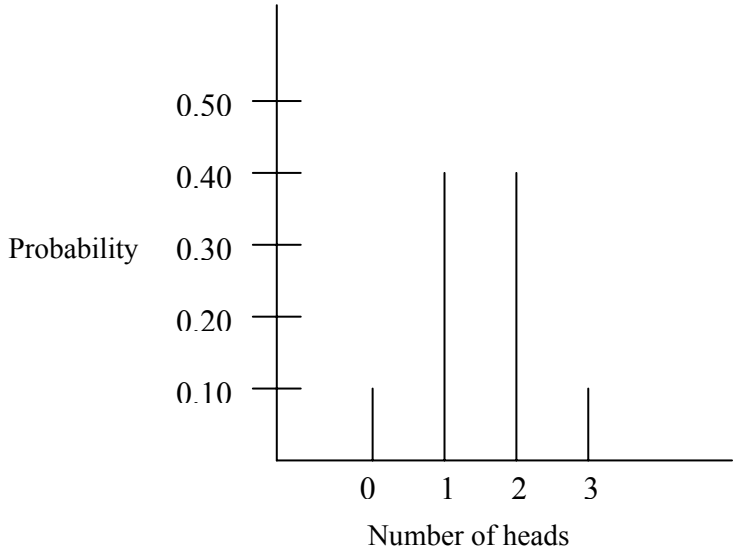


Figure 2: Graphical representation of the distribution of *X*

Graphically along the horizontal axis, plot the various possible values a_i of a random variable and on each value erect a vertical line with height proportional to the corresponding probability p_i.

Now let us consider the c.d.f of the random variable X. Note that if b<0, clearly F(b)=P($X \le b$)=0, because X takes values only {0,1,2,3}. If b=0, that is F(0)=P($X \le 0$)=P(X=0)=1/8. If $0 \le b \le 1$, then P($X \le b$)=P(X=0)+P(0<X<b)=1/8+0=1/8. Similarly, if b=1, F(1)= P($X \le 1$) = P(X=0) + P(X=1)=1/8+3/8=4/8 and so on. Therefore, the c.d.f. F(.) has the following form;

$$F(b)=\begin{cases} 0 & if \quad b<0 \\ \frac{1}{8} & if \quad 0 \le b<1 \\ \frac{4}{8} & if \quad 1 \le b<2 \\ \frac{7}{8} & if \quad 2 \le b<3 \\ 1 & if \quad b \le 3 \end{cases}$$

Note :Mathematical expectation or Expected values or Expectations forms the fundamental idea in the study of probability distribution of any discrete random variable X , the expected value (or mean) , denoted as E(X) is defined as

$$E(X) = x_0p_0+ x_1p_1+ x_2p_2+\ldots\ldots\ldots\ldots=\Sigma x_ip_i$$

Where x_0, x_1, x_2 etc are the values assumed by X and p_0, p_1, p_2 etc are probabilities of these values. Under special conditions (like all probabilities are equal)then

$$E(X) = \text{mean of } x_0, x_1, x_2\ldots\ldots$$

Similarly for continuous variables X having density function p(x) where P[X=x] = p(x) , the Expectation E(X) will be given by integral of $x_ip(x_i)$ w.r.t x

This concept of Expectation also contributes to the definition of **Moment Generating Function of *X i.e* $M_x(t)= E(e^{tx})$**

Example 3 A box contains twice as many red marbles as green marbles. One marble is drawn at random from the box and is replaced ; then a second marble is drawn at random from the box.If both marbles are green you win Rs. 50 ; if both marbles are red you loose Rs. 10 ; and if the y are of different colour then neither you loose nor you win. Determine the probability distribution for the amount you win or loose?

Solution Say X denote the amount you win (+) or loose (-) ; i.e X= +50 or -10

The probability that both marbles are green is 1/9 i.e. P[X=+50] =1/9
The probability that both marbles are red is 4/9 i.e. P[X=-10] =4/9
The probability that marbles are of different colours is 4/9 i.e. P[X=0] =4/9
Thus the probability distribution is given by following table

Amount(in Rs won(+) or lost(-))	**Probability**
+50	1/9
0	4/9
-10	4/9

Check Your Progress 2

E1: Which of the variables given below are discrete? Give reasons for your answer.

(a) The daily measurement of snowfall at Shimla.
(b) The number of industrial accidents in each month in West Bengal.
(c) The number of defective goods in a shipment of goods from a manufacturer.

1.3.1 Binomial Distribution

One very important discrete random variable (or discrete distribution) is the binomial distribution. In this subsection, we shall discuss this random variable and its probability distribution.

Quite often we have to deal with the experiments where there are only two possible outcomes. For example, when a coin is tossed either a head or a tail will comes up, a newborn is either a girl or a boy, a seed either germinates or fails to germinate. Let us consider such an experiment. For example consider the same experiment of tossing a coin independently three times. Note that the coin need not necessarily be a fair one, that is P(Head) may not be equal at P(Tail)

This particular experiment has a certain characteristic. First of all it involves repetition of three identical experiments (trials). Each trial has only two possible outcomes: a Head or a Tail. We refer to the outcome ‘Head’ as success and the outcome ‘Tail’ as failure. All trials are independent of each other. We also know that the probability of getting a ‘Head’ in a trial is p and probability of getting a ‘tail’ in a trial is 1 – p, that is

P(Head) = P(success) = p and P(Tail) = P(failure) = q = 1 - p

This shows that the probability of getting a ‘success’ or a ‘failure’ does not change from one trial to another. If X denotes the total number of ‘Heads’, obtained in three trials, then

X is a random variable, which takes values {0,1,2,3}. Then regarding the above experiment, we have observed the following;

(1) It involves a repetition of n identical trials (Here n=3).
(2) The trials are independent of each other.
(3) Each trial has two possible outcomes.
(4) The probability of success (p) and the probability of failure' (q=1-p) remain constant and do not change from trial to trial.

Now let us try to compute the probabilities $P\{X = 0\}$, $P\{X = 1\}$, $P\{X = 2\}$ and $P\{X = 3\}$ in this case. Note that

$P(X=0)$ = P(getting tails in all three trials)
= $P(\{TTT\}) = (1-p)^3 = q^3$.

Similarly,

$P(X = 1)$ = P(getting one Tail and two Heads in three trials)
= $P(\{THH,HTH,HHT\}) = P(\{THH\}) + P(\{HTH\}) + (\{HHT\})$
= $(1-p)^2p+(1-p)^2p +(1-p)^2p = 3(1-p)^2p = 3q^2p$.

Similarly,

$P(X = 2)$ = (getting two Tails and one Head in three trials)
= $P(\{HTT,THT,TTH\}) = P(\{HTT\}) + P(\{THT\}) + (\{TTH\})$
= $(1-p)p^2 + (1-p)p^2 + (1-p)p^2 = 3(1-p)p^2=3q^2p$

Finally

$P(X=3)$ = P(getting Heads in three trials)
= $P(\{HHH\}) = p^3$

Now observe that instead of n=3, in the above example we can easily compute the probability for any general n. Suppose we compute $P(X = r)$, for $0 \le r \le n$, then note that

$$P(X = r) = C(n,r)p^r(1-p)^{n-r} = C(n,r)p^rq^{n-r},$$

Where C(n,r) denotes the number of ways n places can be filled with r Heads and n-r Tails. From your school mathematics, recall that it is the number of combination of n objects taken r at a time and it can be calculated by the following formula:

$$C(n,r) = \frac{n!}{r!(n-r)!}$$

Therefore, for r = 0,1,…,n,

$$P(X = r) = \frac{n!}{r!(n-r)!} p^r q^{n-r},$$

where
n = the number of trial made
r = the number of success
p = the probability of success in a trial
q = 1-p = the probability of a failure.

Now we define the binomial distribution formally.

Let X represents the number of successes in the set of n independent identical trials. Then X is a discrete random variable taking values 0,1,…,n. The probability of the event P(X= r) is given by

$$P(X=r)=\frac{n!}{r!(n-r)!}p^r q^{n-r},\ \text{r=0,1,2,…..,n}$$

where n, r, p, q are same as defined before. Such a random variable X is called a binomial random variable and its probability distribution is called the binomial distribution. A Binomial distribution has two parameters n and p

Check Your Progress 3

E1 : A farmer buys a quantity of cabbage seeds from a company that claims that approximately 90% of the seeds will germinate if planted properly. If four seeds are planted, what is the probability that exactly two will germinate?

1.3.2 Poisson Distribution

In this subsection we will introduce another discrete distribution called 'Poisson Distribution'. First we shall describe the different situations where we can apply this Poisson Distribution.

Suppose it is the first hour at a bank in a busy Monday morning , and we are interested in the number of customers who might arrive during that hour, or during a 5-minute or a 10-minute interval in that hour. In statistical terms , we want to find the probabilities for the number of arrivals in a time interval.

To find this probability, we are making some assumptions similar to the binomial distribution.

(a) The average arrival rate at any time, remains the same over the entire first hour.
(b) The number of arrivals in a time interval does not depend on what has happened in the previous time intervals.
(c) It is extremely unlikely that more than one customer will arrive at the same time.

Under those above assumptions, we can find the required the probabilities. Suppose X is the random variable denoting the number of customers that arrive in the first hour, then

$$P(X=i)=\frac{e^{-\lambda}\lambda^i}{i!},\qquad i=0,1,2,3,...$$

Where λ (the Greek letter Lambda) denotes the average arrival rate per hour. For example, suppose we know that average number of customers that arrive in that bank during the first hour is 60 and we want to find what is the chance there will be no more than 3 customers in the first 10 minutes. Since we know that the average arrival rate per hour is 60 , if we denote λ to be the average arrival rate per 10 minutes, then $\lambda=\frac{60\times 10}{60}$ =10. Therefore, we can use the above formula and get

$$P(X=i)=\frac{e^{-10}10^i}{i!},\qquad i=0,1,2,3.....$$

But we want to find the probability that no more than 3 customers will be there in the first ten minutes and that is

$$P(X\le 3)=P(X=0)+P(X=1)+P(X=2)+P(X=3) \qquad (1)$$

$$=e^{-10}+e^{-10}10+\frac{e^{-10}10^2}{2!}+\frac{e^{-10}10^3}{3!} \qquad (2)$$

$$\approx 0.00005 + 0.00050 + 0.00226 + 0.00757 = 0.01038 \quad (3)$$

What does this value 0.01038 indicate? It tells us that if the arrival rates are uniform then there is only 1% chance that less than three customers will be there, or in other words, there is a 99% chance that there will be more then 3 customers in the first 10 minutes.

Similarly if we want to find out the chance that there will be no more than 3 customers in the first 5 minutes, then similarly, as above we can see that in this case $\lambda = \frac{60 \times 5}{60} = 5.$

Therefore, if Y denotes the number of customers presents in the first 5 minutes, then

$$P(Y \le 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) \quad (4)$$

$$= e^{-5} + 5e^{-5} + \frac{e^{-5}5^2}{2!} + \frac{e^{-5}5^3}{3!} \quad (5)$$

$$\approx 0.00674 + 0.03369 + 0.08422 + 0.14037 = 0.26502 \quad (6)$$

From the above two examples it is clear that if we change the time unit (and hence the value of λ), the probabilities will change. The probability mass function (p.m.f) given by

$$p(i) = P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}, \quad i = 0,1,2,3,\ldots$$

represents the Poisson probability distribution. From the series expansion of e^{λ}, it easily follows that

$$\sum_{i=0}^{\infty} P(X = i) = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} = 1$$

as it should be.

One point that should always be kept in mind is that a random variable denoting the number of occurrences in an interval of time will follow a Poisson distribution, if the occurrences have the following characteristics:

(a) The average occurrence rate per unit time is constant.
(b) Occurrence in an interval is independent of what has happened previously.
(c) The chance that more than one occurrence will happen at the same time is negligible.

Now let us look at some situations where we can apply the Poisson distribution. Here is an example

EXAMPLE 4: Calls at a particular call center occur at an average rate of 8 calls per 10 minutes. Suppose that the operator leaves his position for a 5 minute coffee break. What is the chance that exactly one call comes in while the operator is away?

Solution: In this case the conditions (a), (b) and (c) are satisfied. Therefore if X denotes the number of calls during a 5 minute interval, then X is a Poisson random variable with $\lambda = \frac{8 \times 5}{10} = 4$. Therefore,

$$P(X = 1) = \frac{e^{-4}4^1}{1!} = 4e^{-4} \approx 0.073$$

That means the chance is 7.3% that the operator misses exactly one call.

Check your progress 4

E 1: If a bank receives on an average $\lambda = 6$ bad Cheques per day, what is the probability that it will receive 4 bad checks on any given day

1.4 CONTINUOUS RANDOM VARIABLE

So far we have discussed about the discrete random variables in details and we have provided two important discrete distributions namely binomial and Poisson distributions. Now in this section we will be discussing another type of random variables namely continuous random variables.

Let us look at the part (a) of Example 2. Note that we want to find the time of occurrence rather than the number of occurrences. Therefore if the random variable X denotes the time of occurrence of a particular event, then the random variable X can take any value on the positive real line or may be any value on a fixed interval say (A,B). Therefore, the random variable can take uncountably many values**.** This type of a random variable which can take uncountably many values is called a continuous random variable. For a continuous random variable X, the probability that X takes a particular value is always zero, but we can always specify the probability of X of any interval through a probability density function (p.d.f.). The exact details are given below.

DEFINITION: Let X be a continuous random variable which takes values in the interval (A,B). A real valued function $f(x)$: $\mathbf{R} \rightarrow \mathbf{R}$ is called the p.d.f of X, if

(a) $f(x) \geq 0$ and $f(x) = 0$, if $x <$ A or $x >$ B.

(b) $\int_A^B f(x)dx = 1$

(c) $P(c < X < d) = \int_c^d f(x)dx$

Now we shall see how we can use the graph of the p.d.f. of a continuous random variable to study real life problems

Example 5: Suppose the Director of a training program wants to conduct a programme to upgrade the supervisory skills of the production line supervisors. Because the programme is self-administered, supervisors require different number of hours to complete the programme. Based on a past study, it is known that the following p.d.f. shows the distribution of time spent by a candidate to complete the program. From the graph it is clear that the average time

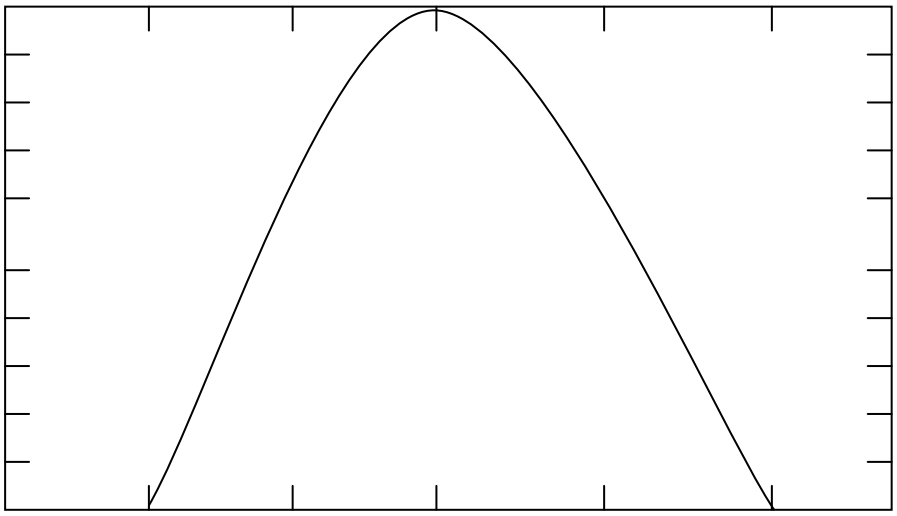


Figure 3: The p.d.f. of the time spent by a candidate to complete the program

Spent by the candidate is 250 and it is symmetrically distributed around 250. How can the Director use this graph to find the following. What is the chance that a participant selected at random will require
(a) more than 250 hours to complete the program
(b) less than 250 hours to complete the program

SOLUTION:Since the graph is symmetric, therefore, it is clear that area under the curve above 250 is half. Therefore, the probability that the random variable takes values higher than 250 is $\frac{1}{2}$. Similarly, the random variable takes value lower than 250 is also $\frac{1}{2}$.
Please try the following exercise now:

Now in the following subsections we will consider different continuous distributions.

1.4.1 The Uniform Random Variable

The uniform distribution is the simplest of a few well-known continuous distributions, which occur quite often. It can be explained intuitively very easily. Suppose X is a continuous random variable such that if we take any subinterval of the sample space, then the probability that X belongs to this subinterval is the same as the probability that X belongs to any other subintervals of the same length. The distribution corresponding to this random variable is called a uniform distribution and this random variable is called a uniform random variable.

Formally, we define the uniform random variable as follows: The random variable X is a uniform random variable between (A, B), if its p.d.f. is given by

$$f(x) = \begin{cases} \dfrac{1}{B-A} & for \quad A < x < B \\ 0 & \text{otherwise} \end{cases}$$

From the p.d.f. it is clear that if A < a_1 < b_1 < B, A < a_2 < b_2 <B and b_1-a_1 = b_2-a_2, then

$P(a_1 < X < b_1) = P(a_2 < X < b_2)$. Therefore, if the length of the intervals are same then the corresponding probabilities will be also equal. Let us see some examples of such random variables:

EXAMPLE 6: A train is likely to arrive at a station at any time uniformly between 6:15 am and 6:30 am. Suppose X denotes the time the train reaches, measured in minutes, after 6:00 am.

SOLUTION In this case X is a uniform random variable takes value between (15,30). Note that in this $P(20 < X < 25)$ is same $P(18 < x < 23)$ and that is equal to

$$\frac{25-30}{30-15} = \frac{23-18}{30-15} = \frac{1}{3}$$

Check your progress 5

E 1 An office fire drill is scheduled for a particular day, and the fire alarm is likely to ring uniformly at any time between 10:00 am to 1:00 pm.

1.4.2 Exponential Random Variable

In making a mathematical model for a real world phenomenon, it is always necessary to make certain simplifying assumptions so as to render the mathematical tractability. On the other hand, however, we can not make too many simplifying assumptions, for then our conclusions obtained from the mathematical model, would not be applicable to the real world problem. Thus in short, we must take enough simplifying assumptions to enable us to handle the mathematics but not so many that the mathematical model no longer resembles the real world problem. One simplifying assumption that is often made is to assume that certain random variables are exponentially distributed. The reason for this is that the exponential distribution is both easy to work and is often a good approximation to the actual distribution.

We use exponential distribution to model lifetime data that is the data, which are mainly non-negative. Although, with proper modifications, it can be used to analyze any type of data (not necessarily non-negative only). The property of the exponential distribution, which makes it easy to analyze, is that it does not deteriorate with time. By this we mean that if the lifetime of an item is exponentially distributed, then an item which has been in use for say ten hours, is as good as a new item in regards to the amount of time remaining until the item fails.

Now we define the exponential distribution formally: A continuous random variable X is said to be an exponential random variable if the p.d.f of X is given for some $X>0$, by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x>0 \\ 0 & \text{otherwise} \end{cases}$$

Here λ is known as the rate constant. It can be shown mathematically that the average value or the mean values of X is $\frac{1}{\lambda}$.Shapes of f (x) for different values of λ are provided in the figure below. From the figure,

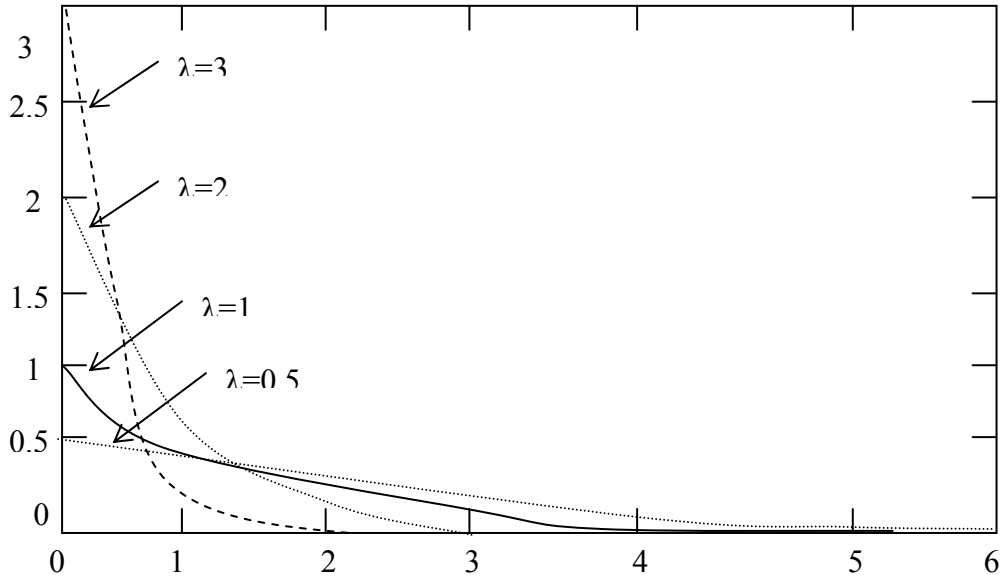


Figure 4: The p.d.f. of the exponential distribution for different values of λ.

It is clear that f(x) is a decreasing function for all values of λ and f(x) tends to 0 as x tends to ∞. Now consider the following example.

EXAMPLE 7: suppose that the amount of time one spends in a bank to withdraw cash from an evening counter is exponentially distributed with mean ten minutes, that is λ = 1/10. What is the probability that the customer will spend more than fifteen minutes in the counter?

SOLUTION: If X represents the amount of time that the customer spend in the counter than we need to find P(*X*>15). Therefore,

$$P(X > 15) = \int_{15}^{\infty} \lambda\, e^{-\lambda x} = e^{-15\lambda} = e^{-\frac{3}{2}} \approx 0.223$$

P(*X*>15)= .223 represents that there is a 22.3 % chance that the customer has to wait more than 15 minutes.

1.4.3 Normal Distribution

Normal distribution is undoubtedly the most used continuous distribution in different areas like astronomy, biology, psychology and of course in probability and statistics also. Because of its practical as well as theoretical importance it has received considerable attentions in different fields. The normal distribution has a unique position in probability theory, and it can be used as an approximation to other distributions. In practice 'normal theory' can frequently be applied with small risk of serious error, when substantially non-normal distributions correspond more closely to the observed value. The work of Gauss in 1809 and 1816 established techniques based on normal distribution, which became standard methods used during the nineteenth century. Because of Gauss's enormous contribution, it is popularly known as Gaussian distribution also.

We will now state the normal distribution formally: The random variable X is said to be normally distributed with parameters μ and σ , if the p.d.f f(x) of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad where \quad -\infty < x < \infty$$

Here μ is a real number lying between -∞ and ∞ and σ is a real number lying between 0 and ∞.

The function f(x) may look a bit strange, but do not get bother. Just notice the following important things. Note that it involves two parameters μ and σ, that means corresponding to each μ and σ we get a distribution function. More over it can be seen that for $-\infty < \mu < \infty$ and $0 < \sigma < \infty$, the function f(x) is symmetric about μ and is a 'bell shaped' one. Both μ and σ have nice interpretation. It can be easily checked that μ is the average value or mean of the distribution and σ provides the measure of spread. The p.d.f. of two different normal distributions are provided below.

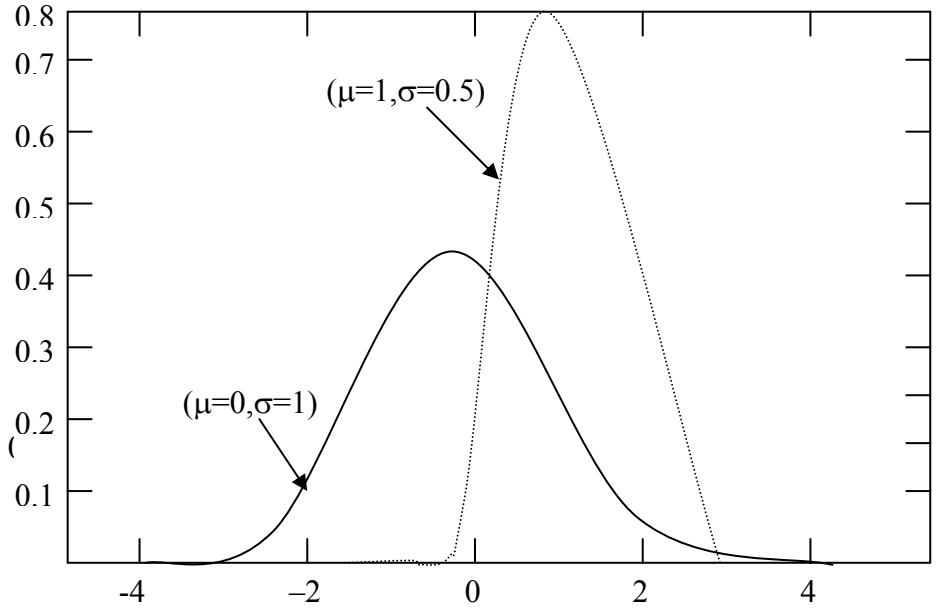


Figure 5: The p.d.f of the normal distribution for two different values of (μ, σ).

It is clear from the above figure that the p.d.f. is symmetric about μ and the shape depends on σ. The spread of the distribution is more if σ is large.

Now let us find the P($a < X < b$) for any a and b, when X follows normal distribution with parameters μ and σ, note that,

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z2}{2}} dz.$$

The last equality follows by simply making the transformation $z = \frac{x-\mu}{\sigma}$. Therefore it follows

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right),$$

Where Z follows normal distribution with parameters 0 and 1. Although the probability can not be calculated in a compact form, extensive tables are available for P (Z < z) for different values of z. The table values can be used to compute P (a < X < b) for any μ and σ.

Say we denote F(a) = P[Z<= a], the probability thet the standard normal variable Z takes values less than or equal to 'a'.The values of F for different values of a are calculated and listed in table . One such table is given in the end of this unit

Note that the entries in the table are values of z for z=0.00 0.01,0.02 , ..0.09.To find the probability that a random variable having standard normal distribution will take on a value between a and b, we use the equation

$$P[a < Z < b] = F(b) - F(a)$$

And if either a or b is negative then we can make use of the identity

$$F(-z) = 1 - F(z)$$

EXAMPLE 8 Use the table to find the following probabilities

(a) P[0.87 < Z < 1.28]
(b) P[-0.34 < Z < 0.62]
(c) P[Z >= 0.85]
(d) P[Z >= -0.65]

SOLUTION

a)P[0.87 < Z < 1.28] : Find F(1.28 from the table). In the table in the row for Z=1.2 find the value under column 0.08 it will be 0.8997 . Similarly find F(0.87) =0.8078

so, P[0.87 < Z < 1.28] = 0.8997 – 0.8078 = 0.0919

b) Similarly P[-0.34 < Z < 0.62] = F(0.62) – F(0.34) = F(0.62) – [1- F(0.34)]
= 0.7324 – (1 - 0.6331) = 0.3655

c) Similarly calculate P[Z > 0.85] = 1- P[Z <= 0.85] = 1-F(0.85) = 0.1977

d)P[Z > -0.65] = 1 – P [Z <= - 0.65]
= 1 – F (-0.65)
= 1- F(1-F(0.65))
= 0.7422

Next we shall see that how to use the standard normal probability table to calculate probability of any normal distribution

Standardising

Any normal random variable X, which has mean Φ and variance σ^2 can be standardized as follows.

Take a variable X, and

i) subtract its mean (m or Φ) and then,
ii) divide by its standard deviation(s or σ).

We will call the result, Z, so

$$Z = \frac{X - \mu}{\sigma}$$

For example, suppose, as earlier, that X is an individual's IQ score and that it has a normal distribution with mean $\Phi = 100$ and standard deviation $\sigma = 15$. To standardize and individuals IQ score, X, we subtract $\Phi = 100$ and divide the result by $\sigma = 15$ to give,

$$Z = \frac{X - 100}{15}$$

In this way every value of X, has a corresponding value of Z. For instance, when X = 130, $Z = \frac{130 - 100}{15} = 2$ and when X = 90, $Z = \frac{90 - 100}{15} = -0.67$.

The distribution of standardized normal random variables

The reason for standardizing a normal random variable in this way is that a standardized normal random variable

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution.

That is, Z is N(0,1). So, if we take any normal random variable, subtract its mean and then divide by its standard deviation, the resulting random variable will have standard normal distribution. We are going to use this fact to calculate (non-standard) normal probabilities.

Calculating probabilities

With reference to the problem of IQ score, suppose we want tot find the probability that an individual's IQ score is less than 85, i.e. P[X<85]. The corresponding area under the pdf N(100,15^2) is shown in Figure 6 below.

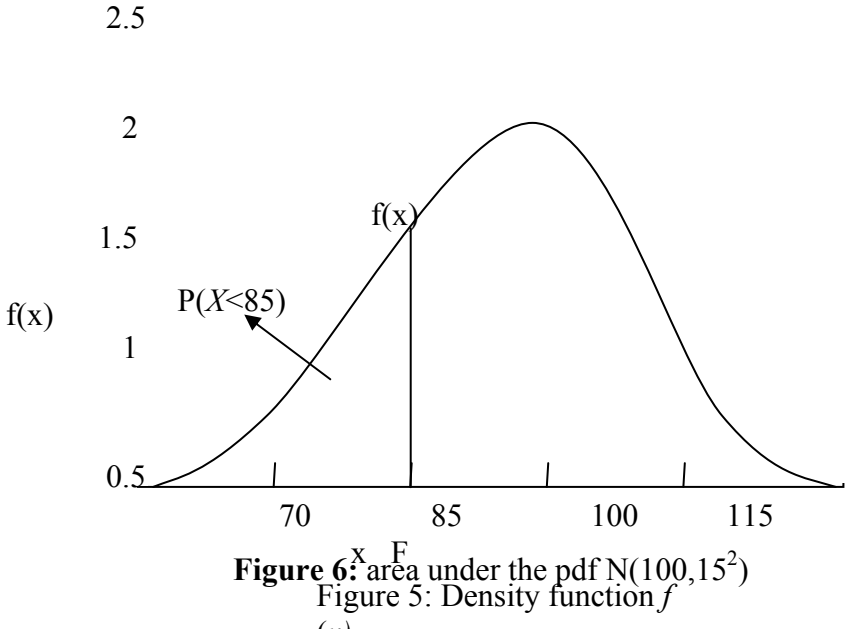


Figure 6: area under the pdf N(100,15^2)

Figure 5: Density function f

We cannot use normal tables directly because these give N(0,1) probabilities. Instead, we will convert the statement X<85 into an equivalent statement which

involves the standardized score, $Z = \frac{X - 100}{15}$ because we know it has a standard normal distribution.
We start with X=85. To turn X into Z we must standardize the X, but to ensure that we preserve the meaning of the statement we must treat the other side of the inequality in exactly the same way. (Otherwise we will end up calculating the probability of another statement, not X<85). 'Standardising' both sides gives,
$\frac{X - 100}{15} < \frac{85 - 100}{15}$.
The left hand side is now a standard normal random variable and so we can call it Z, and we have,

$$Z < \frac{85 - 100}{15}$$

which is

$$Z < -1.$$

So, we have established that the statement we started with, X < 85 is equivalent to Z < – 1. This means that whenever an IQ score, X is less than 85 the corresponding standardized score, Z will be less than – 1 and so the probability we are seeking, P[X<85] is the same P[Z < – 1].

P[Z < – 1] is just a standard normal probability and so we can look it up in Table 1 in the usual way, which gives 0.1587. We get that P[X < 85] = 0.1587.
This process of rewriting a probability statement about X, in terms of Z, is not difficult if you are systematically writing down what you are doing at each stage. We would lay out the working we have just done for P[X < 85] as follows.
X has a normal distribution with mean 100 and standard deviation 15. Let us find the probability that X is less than 85.

$$P[X < 85] = P\left[\frac{X - 100}{15} < \frac{85 - 100}{15}\right]$$

$$= P[Z - 1] = 0.1587$$

Let us do some problems now.

Example 9: For each of these write down the equivalent standard normal probability.

a) The number of people who visit a historical monument in a week is normally distributed with a mean of 10,500 and a standard deviation of 600. Consider the probability that fewer than 9000 people visit in a week.
b) The number of cheques processed by a bank each day is normally distributed with a mean of 30,100 and a standard deviation of 2450. Consider the probability that the bank processes more than 32,000 cheques in a day.

Solution: Here, we want to find the standard normal probability corresponding to the probability P[X < 9000].

a) We have $P[X < 9000] = P\left[\frac{X - 10500}{600} < \frac{9000 - 10500}{600}\right] = P[Z < -2.5]$.

b) Here, we want to find the standard normal probability corresponding to the probability P[X > 32000].

$$P[X < 32000] = P\left[\frac{X - 30100}{2450} < \frac{32000 - 30100}{2450}\right] = P[Z < -0.78]$$

Note: Probabilities like P[a < X < b] can be calculated in the same way. The only difference is that when X is standardized, similar operations must be applied to both a and b. that is, a < X < b becomes,

$$\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}$$

which is

$$\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}$$

Example 10: An individual's IQ score has a N(100, 15^2) distribution. Find the probability that an individual's IQ score is between 91 and 121.

Solution: We require P[91 < X < 121]. Standardising gives

$$P\left[\frac{91-100}{15} < \frac{X-100}{15} < \frac{121-100}{15}\right]$$

The middle term is standardized normal random variable and so we have,

$$P\left[\frac{-9}{15} < Z < \frac{21}{15}\right] = P[-0.6 < Z < 1.4] = 0.9192 - 0.2743 = 0.6449.$$

Check your progress 6

E1 If a random variable has the standard normal distribution , find the probability that it will take on a value

a) Less than 1.50
b) Less than –1.20
c) Greater than –1.75

E2 A filling machine is set to pour 952 ml of oil into bottles. The amount to fill are normally distributed with a mean of 952 ml and a standard deviation of 4 ml. Use the standard normal table to find the probability that the bottle contains oil between 952 and 956 ml ?

1.4.4 Chi-Square Distribution

In the last subsection we have discussed normal distribution. The chi-square distribution can be obtained from the normal distribution as follows. Suppose $Z_1,\ldots,Z_n$ are n independent identically distributed normal random variables with parameters 0 and 1, then $Z_1^2 + \ldots + Z_n^2$ is said to have chi-square distribution with n degrees of freedom. The degrees of freedom here basically indicates the number of independent components which constitute the chi-square distribution. It has received several attention because of its appearance in the constructing analysis of variable tables, contingency tables and for obtaining the critical values of different testing procedure. Now we formally provide the p.d.f of a chi-square random variable with n degrees of freedom.

If the random variable X has chi-square distribution with n-degrees of freedom, then the p.d.f. of X is $f(x)$

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

here Γ (.) is a gamma function and it is defined as

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

Although, the p.d.f of chi-square random variable is not a very nice looking one, do not bother about that. Keep in mind that the shapes of density functions are always skewed. In this case also if we want to compute P(a < X < b) for any a, b and n, explicitly it is not possible to compute. Numerical integration is needed to compute this probability. Even for chi-square distribution extensive tables of P(a < X < b) are available for different values of a, b and n.

Note : We have a standard table corresponding to Chi- Square Distribution, many times you may need to refer the values from the table. So the same is given at the end , method of usage is similar to that discussed in Normal distribution.

EXAMPLE 9 Show that the moment generating function of a random variable X which is chi-square distributed with v degrees of freedom is $M(t) = (1 - 2t)^{-v/2}$.

SOLUTION $$M(t) = E(e^{tx}) = \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^\infty e^{tx} x^{(v-2)/2} e^{-x/2} dx$$

$$= \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^\infty x^{(v-2)/2} e^{-x(1-2t)x/2} dx$$

Letting $(1-2t)^{x/2} = u$ in the last integral we find

$$M(t) = \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^\infty \left(\frac{2u}{1-2t}\right)^{(v-2)/2} e^{-u} \frac{2du}{1-2t}$$

$$= \frac{(1-2t)^{-v/2}}{\Gamma(v/2)} \int_0^\infty u^{(v/2)-1} e^{-u} du = (1-2t)^{-v/2}$$

Check your progress 7

E1 Let X_1 and X_2 be independent random variables, which are chi-square distributed with v_1 and v_2 degrees of freedom respectively. Show that the moment generating function of $Z = X_1 + X_2$ is $(1-2t)^{-v(v_1+v_2)/2}$

E2 Find the values of x^2 for which the area of the right-hand tail of the x^2 distribution is 0.05, if the number of degrees of freedom v is equal to (a) 15, (b) 21, (c) 50.

1.5 SUMMARY

In this unit we have covered following points:

a) A random variable is a variable that takes different values according to the chance outcome

b) Types of random variables: Discrete and Continuous

c) Probability distribution gives the probabilities with which the random variables takes various values in their range

d) Discussed probability distributions :

a. Binomial Distribution : The probability of an event P[X=r] in this distribution is given by

$$P(X = r) = C(n,r)p^r(1-p)^{n-r} = C(n,r)p^r q^{n-r},$$

b. Poisson Distribution : The probability of an event P[X=i] in this distribution is given by

$$P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}, \qquad i = 0,1,2,3,...$$

c. Uniform Distribution : The probability density function is defined by

$$f(x) = \begin{cases} \frac{1}{B-A} & for \quad A < x < B \\ 0 & \text{otherwise} \end{cases}$$

d. Normal Distribution : The probability for this distribution is determined by calculating the area under the curve of probability density function defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad where \quad -\infty < x < \infty$$

e. Chi-Square Distribution : If the random variable X has chi-square distribution with n-degrees of freedom, then the probability density function of X is given by

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1} & \text{if } x > 0 \\ 0 & otherwise \end{cases}$$

here Γ (.) is a gamma function and it is defined as

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

f. Mathematical expectation or Expected values or Expectations E(X) is defined as E(X) = $x_0p_0+ x_1p_1+ x_2p_2+$……..=Σx_ip_i when all probabilities are equal then E(X) = mean of x_0, x_1, x_2……
Similarly for continuous variables X having density function p(x) where P[X=x] = p(x) , the Expectation E(X) will be given by integral of $x_ip(x_i)$ w.r.t x .
This concept of Expectation also contributes to the definition of Moment Generating Function of X *i.e* $M_x(t)= E(e^{tx})$

1.6 SOLUTIONS

Check Your Progress 1

E1 a) If X denotes the number of correct answers, then X is the random variable for this situation

b) X can take the values 0,1,2,3…..up to 50

c) P[$X=10$] means the probability that the number of correct answers is 10

Check Your Progress 2

E1 Case (a) is not discrete where as case (b) and (c) are discrete because in case (a) we are taking values in an interval but in case(b) the number of accident is finite , similarly you argue for case (c)

Check Your Progress 3

E1 This situation follows the binomial distribution with n=4 and p=90/100=9/10 The random variable X is the number of seeds that germinate. We have to calculate the probability that exactly two of the four seedswill germinate. That is P[X=2]. By applying Binomial formula , we get

$P[X=2] = ({}^4C_2) * (9/10)^2 * (1/10)^2$

$= 6 * (81/100) * (1/100) = 486/10000 = 0.0486$

So , the required probability is 0.0486

Check Your Progress 4

E1 Here we are dealing with the problem related to the receipt of bad Cheques, which is an event with rare occurrence over an interval of time (which is a day In this case). So, we can apply Poisson distribution

Average bad Cheques received per day = 6
Thus by substituting $\lambda = 6$ and x=4 in Poisson formula we get

$$P[X=4] = (6^4 e^{-6})/4! = 0.135$$

Check your progress 5

E1 Suppose X denotes the time the fire alarm starts measured in minutes after 10:00 am. Then clearly X is a uniform random variable between (0,180). If we want to find the probability that fire alarm will ring before noon, then

$$P(X \le 12:00 \text{ noon}) = \frac{(12-10)\times 60}{180} = \frac{2}{3}.$$

Check your progress 6

E1 a) 0.9332
b) 0.1151
c) 0.9599

E2 The standard normal probability corresponding to this probability is given by

$$\begin{aligned} P[952 < Z < 956] &= P[((952-952)/4) < ((X-952)/4) < ((952-956)/4)] \\ &= P[0 < Z < 1] \\ &= F(1) - F(0) \\ &= 0.8413 - 0.5 = 0.343 \end{aligned}$$

Check your progress 6

E1 The moment generating function of $Z = X_1 + X_2$ is

$$M(t) = E\left[e^{t(X_1+X_2)}\right] = E(e^{tX_1})E(e^{tX_2}) = (1-2t)^{-v_1/2}(1-2t)^{-v_2/2} = (1-2t)^{-(v_1+v_2)/2}$$

using **Example 9**.

E2 Using the table in for Chi Square distribution we find in the column headed $x^2_{.95}$ the values: (a) 25.0 corresponding to $v = 15$; (b) 32.7 corresponding to $v = 21$; (c) 67.5 corresponding to $v = 50$.

UNIT 2 PSEUDO RANDOM NUMBER GENERATION

2.0 INTRODUCTION

A pseudo-random number generation is the methodology to develop algorithms and programs that can be used in, probability and statistics applications when large quantities of random digits are needed. Most of these programs produce endless strings of single-digit numbers, usually in base 10, known as the decimal system. When large samples of pseudo-random numbers are taken, each of the 10 digits in the set {0,1,2,3,4,5,6,7,8,9} occurs with equal frequency, even though they are not evenly distributed in the sequence.

Many algorithms have been developed in an attempt to produce truly random sequences of numbers, endless strings of digits in which it is theoretically impossible to predict the next digit in the sequence based on the digits up to a given point. But the very existence of the algorithm, no matter how sophisticated, means that the next digit can be predicted! This has given rise to the term pseudo-random for such machine-generated strings of digits. They are equivalent to random-number sequences for most applications, but they are not truly random according to the rigorous definition.

A simulation that has any random aspects at all, must involve sampling or generating random variables from different probability distributions. These distributions are often specified, that is the form of the distribution functions is explicitly known, for example it can be exponential, gamma, normal or Poisson as discussed in Unit 1.

Random number generation has intrigued scientists for several years and a lot of efforts has been spent on the creation of randomness on a deterministic (non-random) machine, that is to design computer algorithms that are able to produce 'random' sequences of integers. This is not a trivial task. Such algorithms are called generators and all generators have flaws because all of them construct the n-th number in the sequence as a function of the $(n-1)$ – th number, initialized with a non-random seed value. Numerous techniques have been invented over the years that measure just how random a sequence is, and most well known generator have subjected to rigorous testing. The mathematical tools that are required to design such an algorithm are largely number theoretic and combinatorial in nature. These tools differ drastically from those needed when we want to generate sequences of integers with certain non-uniform distributions given that a perfect uniform random number generator is available.

The methodology of generating random numbers has a long and interesting history. The earliest methods were essentially carried out by hand, such as casting lots, throwing dice, dealing out cards or drawing numbered balls from a well-stirred urn. Many lotteries are still operating in this way. In the early twentieth century statisticians joined gamblers in generating random numbers and mechanized devices were built to generate random numbers more quickly. Some time later, electric circuits based on randomly pulsating vacuum tubes were developed that delivered random digits at rates up to 50 per second. One such random number generator machine the Electronic Random Number Indicator Equipment (ERNIE) was used by the British General Post Office to pick the winners in the Premium Savings Bond lottery. Another electronic device was used by the Rand Corporation to generate a table of million random digits. In India also Indian Statistical Institute has published a book just the collection of million random numbers in the mid twentieth century which was used for sample survey planning.

As computers and simulation became more widely used, increasing attention was paid to methods of random number generation compatible with the way computers work. A good uniform between 0 and 1, random generator should possess certain properties as listed below:

- Above all, the numbers produced should appear to be distributed uniformly on [0, 1] and should not exhibit any correlation with each other; otherwise, the simulation's results may be completely invalid.
- From a practical point of view a generator should be fast and avoid the need for a lot of storage.
- We should be able to produce a given stream of random numbers from a given initial (seed) value for at least two reasons. First this can sometimes make debugging or verification of the computer program easier or we might want to use identical random numbers in simulating different systems in order to obtain a more precise comparison.

In this unit we will describe how to generate $U(0,1)$ (uniform between 0 and 1, see unit 1 for the actual definition) in a

computer. Once we have a U (0, 1) random number, we will use that to generate several other deviates from different discrete and continuous distributions.

2. 1 OBJECTIVES

After reading this unit, you should know

- how to generate U (0, 1) random number in a computer
- how to generate random deviates from any discrete distribution
- how to generate random numbers from many continuous distributions, like exponential, Weibull, gamma, normal, chi-square etc.

2.2 UNIFORM RANDOM NUMBER GENERATORS

As we had mentioned in the previous section that we need the uniform random number of generator for generating random numbers from any other distributions. Therefore, it is very important to have a very good uniform random number generator. There are several methods available to generate uniform random numbers. But currently the most popular one is the linear congruential generator (LCG). Most of the existing software's today use this LCGs proposed by Lehmer in the early 50's. The LCGs provides an algorithm how to generate uniform random number between (0,1). It can be simply described as follows.

A sequence of integers $Z_1, Z_2,\ldots$ is defined by the recursive formula

$$Z_i = (aZ_{i-1} + c) \text{ (mod m)}, \qquad (1)$$

where m, the modulus, a, the multiplier, c, the increment and Z_0, the seed or the initial value, are all non-negative integers. Those who are not familiar with the definition of modules, note that for non-negative integers, x, y and z, $x = y$ (mod z) means x is the remainder when the integer y is divided the integer z. For example if $y = 10$ and $z = 3$, then $x = 1$, or if $y = 10$ and $z = 2$, then $x = 0$. Therefore, from (1) it is clear that to obtain Z_i, first divide $aZ_{i-1} + c$ by m and Z_i is the corresponding remainder of this division. It is clear that $1 \le Z_i \le m - 1$ and to obtain the desired random numbers U_i, for $i = 1, 2\ldots$. On [0, 1], we let $U_i = \frac{Zi}{m}$. The choice of the non-negative integers, a, c and m are the most crucial steps, and they come from the theoretical considerations. In addition to non-negatively, they also should satisfy $0 < m$, $a < m$ and $c < m$. Moreover, the initial value $Z_0 < m$.

Immediately, two objections could be raised against LCGs. The first objection is one which is common to all random number generators, namely the Z_i's defined by (1) are not really random. It can be easily seen that for $i = 1, 2,\ldots.$,

$$Zi = \left[a^i Z_0 + \frac{c(a^i - 1)}{a - 1} \right] \text{(mod m)},$$

so that every Z_i is completely determined by m, a, c and Z_0. However, by careful choice of these four parameters the aim is to induce a behavior in the Z_i's that makes the corresponding U_i's appear to be independent identically distributed U (0, 1) random variates when subjected to variety to statistical tests.

The second objection to LCGs might be that the U_i's can take on only the rational numbers 0, $\frac{1}{m}$, $\frac{2}{m}$, …., $\frac{(m-1)}{m}$; in fact the U_i's might take only a fraction of these values, depending on the specifications of the four parameters. Therefore, there is no possibility of getting a value of U_i between, say $\frac{0.1}{m}$ and $\frac{0.9}{m}$, whereas this should occur with probability $\frac{0.8}{m} > 0$. We will see later that the modulus m is usually chosen to be very large, say 10^9 or more, so that the points in [0, 1] where U_i's can fall are very dense. This provides an accurate approximation to the true continuous U (0, 1) distribution, sufficient for most purposes.

Let us consider the following example:

Example 1 : Consider the LCG defined by $m = 16$, $a = 5$, $c = 3$ and $Z_0 = 7$. The following table gives Z_i and U_i (up to three decimal places) for $i = 1$, …., 19. Note that $Z_{17} = Z_1 = 6$, $Z_{18} = Z_2 = 1$ and so on. Therefore, from $i = 17$, the sequence repeats itself. Naturally we do not seriously suggest anybody to use this generator. The main reason here m is too small. This we are presenting just for illustrative purpose.

Table 1

The LCG $Z_i = (5Z_{i-1} + 3)$ (mod 16) with $Z_0 = 7$

i	Z_i	U_i	i	Z_i	U_i	i	Z_i	U_i	i	Z_i	U_i

0	7	-	5	10	0.625	10	9	0.563	15	4	0.250
1	6	0.375	6	5	0.313	11	0	0.000	16	7	0.438
2	1	0.063	7	12	0.750	12	3	0.188	17	6	0.375
3	8	0.500	8	15	0.938	13	2	0.125	18	1	0.063
4	11	0.688	9	14	0.875	14	13	0.813	19	8	0.500

Note that the repeating behaviour of LCG is inevitable. By the definition of Z_i, whenever it takes on a value it had taken previously, from that point onward the sequence will repeat itself endlessly. The length of a cycle is called the period of a generator. For LCG, Z_i depends only on the previous value Z_{i-1} and since $0 \le Z_i \le m - 1$, it is clear that the period is at most m. If it m, the LCG is said to have full period. Clearly, if a generator is full period, any choice of the initial seed Z_0 from $\{0,\ldots,m\text{-}1\}$ will produce the entire cycle in some order.

Since for large scale simulation projects may require hundreds of thousands of random numbers, it is desirable to have LCGs with long periods. Moreover, it is desirable to have full period LCGs, because then it is assured to have every integer between 0 and m – 1 exactly once in every cycle. Thus it is very important to know how to choose a, m and c so that the corresponding LCG will have full period. We should also keep in mind that obtaining full period is just one desirable property for a good LCG. It should also have good statistical properties, such as apparent independence, computational and storage efficiency and reproducibility. Reproducibility is simple. For reproducibility, we must only remember that the initial seed Z_0 initiates the generator with this value again to obtain the same sequence of U_i exactly. Some of the standard LCGs with different values of a, m and c are presented below. These LCG have been observed to perform very well in several machines and passed the standard statistical tests also.

Generator 1: $a = 16807$, $m = 2^{31} - 1$, $c = 0$.
Generator 2: $a = 1664525$, $m = 2^{32}$, $c = 1664525$.

Fortunately today, most of the simulation packages and even simple scientific calculators have reliable U (0 1,) generator available.

Check your progress 1

E1 What do you mean by Pseudo random number generation? What is the practical advantage of the concept of random number generation? Do you know any algorithm which works in designing the software for Random number generation?

2.3 GENERATING RANDOM VARIATES FROM ARBITRARY DISTRIBUTIONS

A simulation that has any random aspects at all must involve generating random variates from different distributions. We usually use the phrase generating a random variate to refer the activity of obtaining an observation or a realization on a random variable from the desired distribution. These distributions are often specified as a result of fitting some appropriate distributional form. They are often specified in advance, for example exponential, gamma or Poisson etc. In this section we assume that the distributional form is already been specified including the values of the parameters and we address the issue of how we can generate random variate from this specified distribution.

We will see in this section that the basic ingredient needed for every method of generating random variates from any distribution is a source of independent identically distributed U (0, 1) random variates. Therefore, it is essential that a statistically reliable U (0, 1) generator is available for generating random deviate correctly from any other distribution.Therefore, from now on we assume that we have a reliable sequence of U(0,1) variates available to us.

There are several issues which should be kept in mind before using any particular generator. The most important issue is of course the exactness. It is important that one should use an algorithm that results in random variates with exactly the desired distribution, within the unavoidable external limitations of machine accuracy and the exactness of U (0,1) random number generator. The second important issue is efficiency. Given that we have several choices, we should choose that algorithm which is efficient in terms of both storage space and execution time. Now we provide some of the most popular and standard techniques to generator non-uniform random deviates, it may be both discrete or continuous and even mixed distribution also.

2.4 INVERSE TRANSFORM

Suppose we want to generate a random variate X that has a continuous and strictly increasing distribution function F, when $0 < F(x) < 1$, i.e., whenever $x_1 < x_2$ then $F(x_1) < F(x_2)$. We draw a curve of F in the Figure 1. Let F^{-1} denote the inverse of the function F. Then an algorithm for generating a random variate X having distribution function F is as follow .

ALGORITHM

- Step 1: Generate U_1 from U_1 (0,1)
- Step 2: Return $X_1 = F^{-1}(U_1)$.

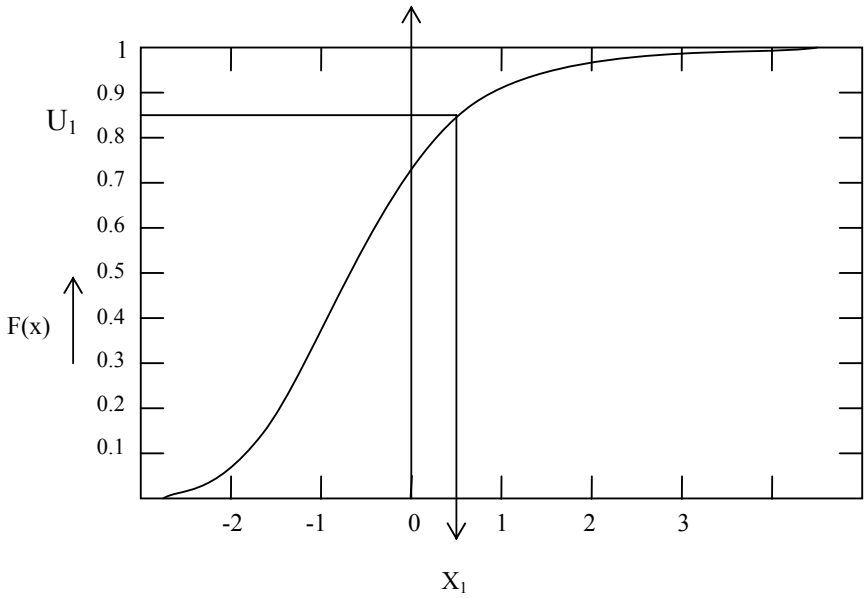


Figure 1: Distribution function of a continuous a random variable.

Note that when F is a strictly increasing function, $F^{-1}(U)$ will be always defined, since $0 \le U \le 1$ and the range of F is [0, 1]. Figure 1 illustrates the algorithm graphically. According to the figure it is clear that the uniform random variable U_1 results the random variable X_1 with the distribution function F. To show that the value X_1 returned by the above algorithm has the desired distribution function F, note that

$$P(X_1 \le x) = P(F^{-1}(U_1) \le x) = P(U_1 \le F(x)) = F(x).$$

The first equality follows from the definition of X_1, the second equality follows because F is invertible and the third equality follows because U_1 follows $U(0,1)$.

EXAMPLE 2: Let X have the Weibull distribution with the following probability density function:

$$f(x) = \begin{cases} \alpha\lambda e^{-\lambda x^{\alpha}} x^{\alpha-1} & if \quad x > 0 \\ 0 & if \quad x < 0 \end{cases}$$ Find F^{-1}

SOLUTION Here α and λ both are known constants and both of them are strictly greater than 0.
Therefore, X has the distribution function

$$F(x) = \begin{cases} 1 - e^{-\lambda x^{\alpha}} & if \quad x > 0 \\ 0 & if \quad x \le 0 \end{cases}$$

Therefore, to compute $F^{-1}(u)$, let us equate $u = F(x)$ and we solve for x to obtain

$$F^{-1}(u) = \left[\frac{1}{\lambda}\ \{-\ln(1-u)\}\right]^{\frac{1}{\alpha}}$$

Therefore to generate X from a Weibull distribution with $\alpha = 2$ and $\lambda = 1$, generate U from $U(0, 1)$ and then set

$$X = [\{-\ln(1-u)\}]^{\frac{1}{2}}$$

In this case also as before, it is possible to replace U by $1 - U$, therefore we can use

$$X = [\{-\ln u\}]^{\frac{1}{2}},$$

to avoid one subtraction.

The inverse-transform method can be used also when the random variable X is discrete.
In this case the distribution function is

$$F(x) = P(X \le x) = \sum_{x_i \le x} p(x_i),$$

Where $p(x_i)$ is the probability mass function of X, *i.e.,*

$$p(x_i) = P(X = x_i).$$

We can assume that X can take values $x_1, x_2, \ldots$, where $x_1 < x_2 < \ldots$. Then the algorithm is as follows:

ALGORITHM

- Step 1: Generate U from $U(0,1)$
- Step 2: Determine the smallest positive integer I such that $U \le F(x_1)$ and return $X = x_1$. The following figure 2 illustrates the method. In that case we generate $X=x_4$

Now to show that the discrete inverse transform method is valid, we need to show that $P(X = x_i) = p(x_i)$ for all $i = 1$, we get $X = x_1$, if and only if $U \le F(x_1) = p(x_1)$, since x_i's are in the increasing order and U follows $U(0,1)$. For $i \ge 2$, the algorithm sets $X = x_i$ if and only if $F(x_{i-1}) < U \le F(x_i)$, since the i chosen by the algorithm is the smallest positive integer such that $U \le F(x_i)$. Further, since U follows $U(0,1)$ and $0 \le F(x_{i-1}) < F(x_i) \le 1$,

$$P(X = x_i) = P\{F(x_{i-1} < U \le F(x_i)\} = F(x_i) - F(x_{i-1}) = p(x_i).$$

Now consider the following example.

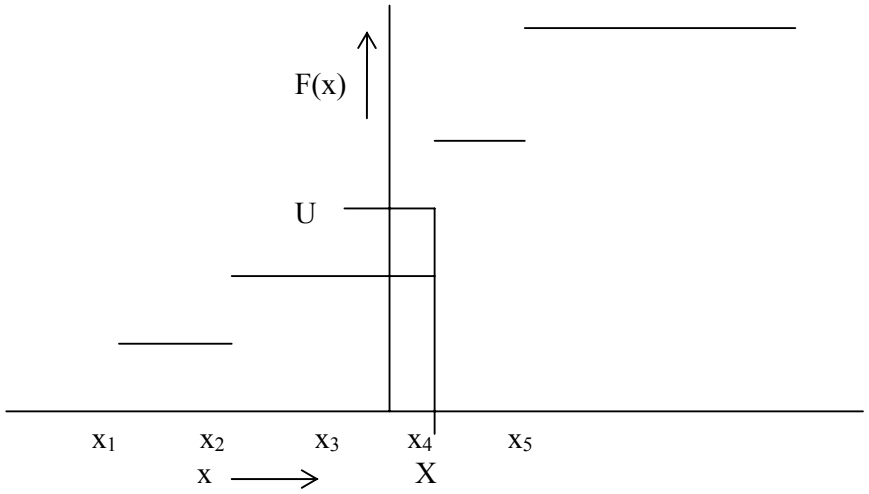


Figure 2: Distribution function of a discrete random variable

EXAMPLE 3: Suppose we want to generate a random sample from the following discrete probability distribution.

$P(X = 1) = \frac{1}{2}$, $P(X = 2) = \frac{1}{4}$, $P(X = 4) = \frac{1}{4}$. Generate a random sample from X?

SOLUTION The distribution function of the random variable X is

$$F(x) = \begin{cases} 0 & if \quad x < 1 \\ \frac{1}{2} & if \quad 1 \le x < 2 \\ \frac{3}{4} & if \quad 2 \le x < 4 \\ 1 & if \quad x \ge 4 \end{cases}$$

The distribution function is X is presented in the Figure 3. If we want to generate a random sample from X, first generate a random variable U from $U(0, 1)$. If $U \le \frac{1}{2}$, then assign $X = 1$. If $\frac{1}{2} < U \le \frac{3}{4}$, then assign $X = 2$, otherwise assign $X = 4$.

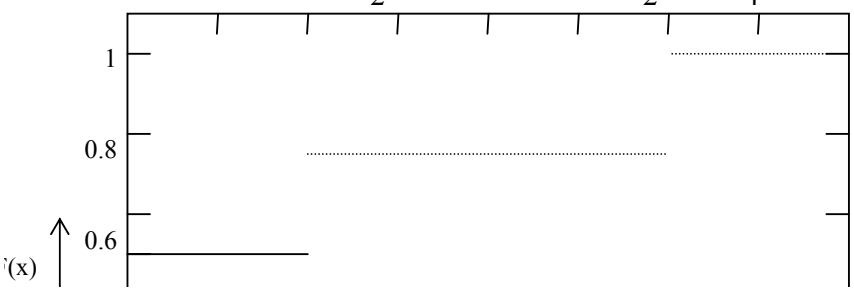


Figure 3: Distribution function of the random variable X of example 3.

Check your progress 2

E1 Let X have the exponential distribution with mean 1. The distribution function is $F(x) = \begin{cases} 1-e^{-x} & if \quad x \geq 0 \\ 0 & if \quad x < 0 \end{cases}$ Find F^{-1}

E2 Consider another discrete random variable which may take infinitely many values. Suppose the random variable X has the following probability mass function.

$P(X = i) = p_i = \frac{1}{2^i}$, i = 1, 2, 3, …… . Generate a random sample from *X?*

2. 5 ACCEPTANCE-REJECTION METHOD

In the last subsection we have discussed the inverse transformation method to generate random number from different non-uniform distributions. Note that apparently, the inverse transformation method seems to be the most general method to generate random deviate from any distribution function functions. In fact it can be used provided the distribution function can be written in an explicit form, or more precisely the inverse of the distribution function can be computed analytically. For example, in case of exponential, Weibull, Cauchy distributions the distribution function and also their inverses can be constructed analytically. Unfortunately that is not the case in general. Suppose the random variable X follows gamma with the shape and scale parameters as α and λ respectively. Then the density function of X, say
$fx(x \mid \alpha, \lambda)$, is

$$f(x \mid \alpha, \lambda) = \begin{cases} \frac{\lambda^{\alpha}}{\Gamma\alpha} x^{\alpha-1} e^{-\lambda x} & if \quad x > 0 \\ 0 & if \quad x \leq 0 \end{cases}$$

The distribution function X, say $F(x \mid \alpha, \lambda) = \int_0^x f(y \mid \alpha, \lambda) dy$ can not be expressed in explicit form. Therefore, $F^{-1}(x \mid \alpha, \lambda)$ also can not be calculated explicitly. Exactly the same problem arises if X is a normal random variable. Suppose X is a normal random variable with mean μ and variance σ^2, then the probability density function of X, say $f(x \mid \mu, \sigma)$ is

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)2}{2\sigma^2}}, \text{ for } -\infty < x < \infty.$$

In this case also the distribution function can not be computed analytically and similarly it's inverse also. Therefore in these cases we can not apply the inverse transformation method to generate the corresponding random deviates. The acceptance-rejection method can be used quite effectively to generate these random deviates. It can be described as follows.

Suppose we have a density function $f(x)$ and we want to generate a random deviate from the density function $f(x)$. The distribution function of $f(x)$ can not be expressed in explicit form. The acceptance-rejection method requires that we specify a function $g(x)$ that majorizes the function $f(x)$, that is, $f(x) \leq g(x)$ for all x. Naturally, $g(x)$ will not be a density function always, since

$$c = \int_{-\infty}^{\infty} g(x) dx \geq \int_{-\infty}^{\infty} f(x) = 1,$$

but the function $h(x) = \frac{1}{c} g(x)$ is clearly a density function provided $c < \infty$. Now for any given $f(x)$, we choose the function $g(x)$, such that $c < \infty$ and it is possible to generate random deviate from $g(x)$ by a inverse transformation method. Then we can generate the random deviate X from $f(x)$ as follows:

ALGORITHM

- Step 1: Generate Y having density function $g(x)$.
- Step 2: Generate U from $U(0,1)$ which is independent of Y.
- Step 3: If $U \le \dfrac{f(Y)}{g(Y)}$, $X = Y$, otherwise go back to Step 1 and try again.

Note that the algorithm is looping back to Step 1 until we generate a pair (Y, U) pairs in Steps 1 and 2 for which $U \le \dfrac{f(Y)}{g(Y)}$, when we accept the value Y for X. Theoretically it can be shown that the random deviate X generated by the above algorithm has indeed the probability density function $f(x)$. Since it is not very easy to prove the result we do not provide it.

EXAMPLE 4**:** Consider the following density function

$$f(x) = \begin{cases} 60x^3(1-x)^2 & if \quad 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

In this case the distribution function of $f(x)$ is

$$F(x) = \begin{cases} 0 & if \quad x < 0 \\ 10x^6 + 15x^4 - 24x^5 & if \quad 0 < x < 1 \\ 1 & if \quad x > 1 \end{cases}$$

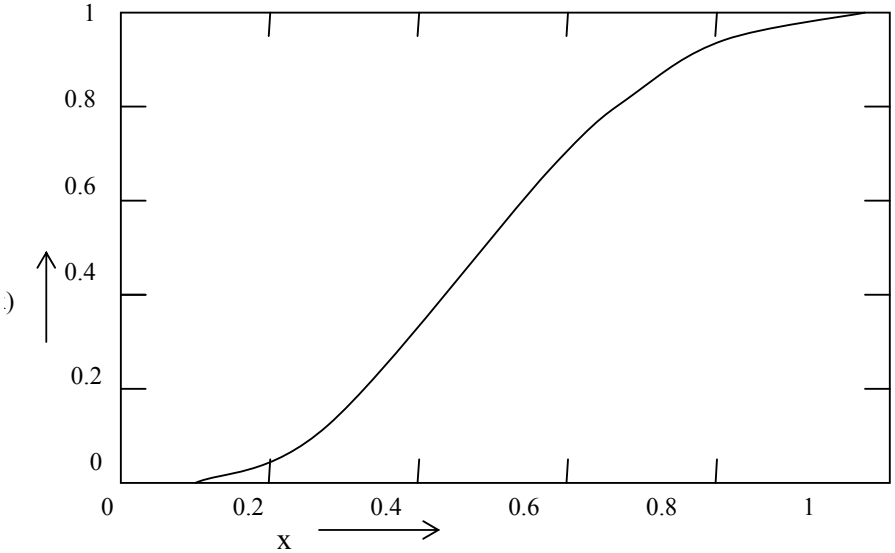
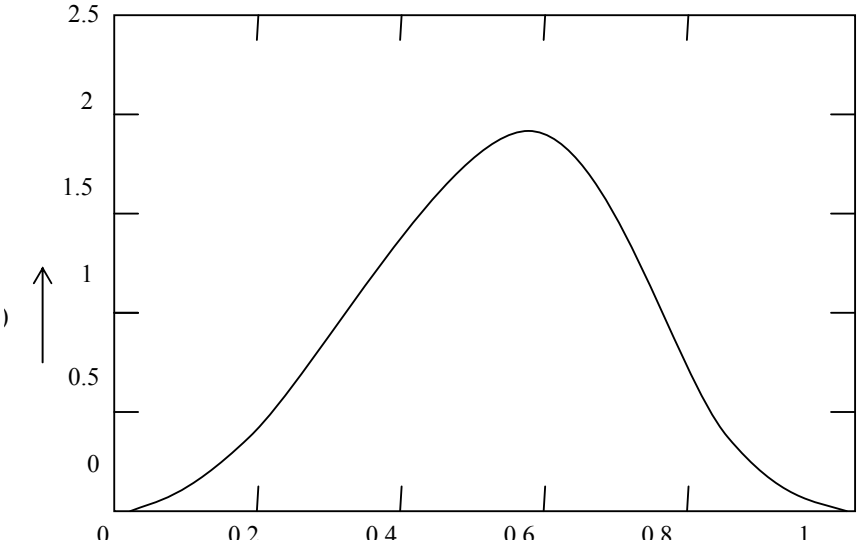


Figure 4 : Distribution function of the density function $f(x)$



From the distribution function $F(x)$ is provided in figure 4 .It is clear that the distribution function $F(x)$ is a strictly increasing function of x in [0, 1]. Therefore $F^{-1}(x)$ exists, but unfortunately to find $F^{-1}(x)$ we need to solve a six degree polynomial, which can not be obtained analytically. We need to solve numerically only. Therefore, we can not generate random deviate from the density function $f(x)$ using the inversion method. But we will be able to generate random deviate from $f(x)$ using the acceptance-rejection method.

First let us look at the graph of the density function $f(x)$. It is presented in the Figure 5. From the Figure 5 it is clear that $f(x)$ is an unimodal function with a unique maximum. The maximum can be easily obtained by the standard differential calculus, that is by setting $\frac{df(x)}{dx} = 0$. We see that the maximum of $f(x)$ occurs at $x = 0.6$ and the maximum value at 0.6, that is $f(0.6) = 2.0736$. Therefore, if we define

$$g(x) = \begin{cases} 2.0736 & if \quad 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

then clearly $f(x) \le g(x)$ for all x. Now to calculate $h(x)$, first we need to calculate c. Since,

$c = \int_0^1 2.0735 = 2.0736$, therefore,

$$h(\underline{x}) = \begin{cases} \frac{2.0736}{c} = 1 & if \quad 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

It is immediate that $h(x)$ is just the U(0,1) density function. Now the algorithm takes the following form.

- Step 1: Generate Y from $U(0, 1)$
- Step 2: Generate U from $U(0,1)$ which is independent of Y.
- Step 3: If $U \le \frac{60Y^3(1-Y)^2}{2.0736}$ then return with $X = Y$,otherwise go back to Step 1.

In this case X has the desired density function $f(x)$.

Check your progress 3

E1 use acceptance-rejection method to generate a random deviate from gamma density function, . The gamma density function with the shape parameter α can be written as

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} & if \quad x > 0 \\ 0 & \text{otherwise} \end{cases}$$

2. 6 SUMMARY

In this unit we have discussed the meaning of pseudo random number generation and along with that we have described uniform random number generator and arbitrary random number generator. Under Uniform random number generator case we have emphasized on **LCG (Linear Congruential Generator)** and some objections related to LCG. Actually LCGs provides an algorithm how to generate uniform random number between (0,1). It can be simply described as follows.

A sequence of integers $Z_1, Z_2, \ldots$ is defined by the recursive formula

$$Z_i = (aZ_{i-1} + c) \text{ (mod m)}, \qquad (1)$$

where m, the modulus, a, the multiplier, c, the increment and Z_0, the seed or the initial value, are all non-negative integers

Then we have discussed the concept , algorithm and application of **Inverse transforms** for random number generation . In brief Suppose we want to generate a random variate X that has a continuous and strictly increasing distribution function F, when $0 < F(x) < 1$, i.e., whenever $x_1 < x_2$ then $F(x_1) < F(x_2)$. Let F^{-1} denote the inverse of the function F. Then an algorithm for generating a random variate of X having distribution function F is as follow .

ALGORITHM

- Step 1: Generate U_1 from U_1 (0,1)
- Step 2: Return $X_1 = F^{-1}(U_1)$.

The inverse-transform method can be used also when the random variable X is discrete.
In this case the distribution function is

$$F(x) = P(X \le x) = \sum_{x_i \le x} p(x_i),$$

Where $p(x_i)$ is the probability mass function of X, *i.e.,*

$$p(x_i) = P(X = x_i).$$

We can assume that X can take values $x_1, x_2, \ldots$, where $x_1 < x_2 < \ldots$. Then the algorithm is as follows:

ALGORITHM

- Step 1: Generate U from U (0,1)
- Step 2: Determine the smallest positive integer I such that $U \le F(x_1)$ and return $X = x_1$. The following figure 2 illustrates the method. In that case we generate $X=x_4$

Note:

- the inverse transformation method to generate random number from different non-uniform distributions. Note that apparently, the inverse transformation method seems to be the most general method to generate random deviate from any distribution function functions. In fact it can be used provided the distribution function can be written in an explicit form

Finally we had discussed the **Acceptance Rejection method** of random number generation, this method is quite important because ,when the distribution function can not be computed analytically and similarly it's inverse also. Then in such cases we can not apply the inverse transformation method to generate the corresponding random deviates. The acceptance-rejection method can be used quite effectively to generate these random deviates.

In brief ,Suppose we have a density function $f(x)$ and we want to generate a random deviate from the density function $f(x)$. The distribution function of $f(x)$ can not be expressed in explicit form. The acceptance-rejection method requires that we specify a function $g(x)$ that majorizes the function $f(x)$, that is, $f(x) \le g(x)$ for all x. Naturally, $g(x)$ will not be a density function always, since

$$c = \int_{-\infty}^{\infty} g(x)dx \ge \int_{-\infty}^{\infty} f(x) = 1,$$

but the function $h(x) = \frac{1}{c} g(x)$ is clearly a density function provided $c < \infty$. Now for any given $f(x)$, we choose the function $g(x)$, such that $c < \infty$ and it is possible to generate random deviate from $g(x)$ by a inverse transformation method. Then we can generate the random deviate X from $f(x)$ as follows:

ALGORITHM

- Step 1: Generate Y having density function $g(x)$.
- Step 2: Generate U from U (0,1) which is independent of Y.
- Step 3: If $U \le \frac{f(Y)}{g(Y)}$, $X = Y$, otherwise go back to Step 1 and try again.

Note that the algorithm is looping back to Step 1 until we generate a pair (Y, U) pairs in Steps 1 and 2 for which $U \le \frac{f(Y)}{g(Y)}$, when we accept the value Y for X. Theoretically it can be shown that the random deviate X generated by the above algorithm has indeed the probability density function $f(x)$. Since it is not very easy to prove the result we do not provide it.

2. 7 SOLUTIONS

Check your progress 1

E1 A pseudo-random number generation is the methodology to develop algorithms and programs that can be used in, probability and statistics applications when large quantities of random digits are needed. Most of these programs produce endless strings of single-digit numbers, usually in base 10, known as the decimal system. When large samples of pseudo-random numbers are taken, each of the 10 digits in the set {0,1,2,3,4,5,6,7,8,9} occurs with equal frequency, even though they are not evenly distributed in the sequence.

Many algorithms have been developed in an attempt to produce truly random sequences of numbers, endless strings of digits in which it is theoretically impossible to predict the next digit in the sequence based on the digits up to a given point. But the very existence of the algorithm, no matter how sophisticated, means that the next digit can be predicted! This has given rise to the term pseudo-random for such machine-generated strings of digits. They are equivalent to random-number sequences for most applications, but they are not truly random according to the rigorous definition.

There are several methods available to generate uniform random numbers. But currently the most popular one is the linear congruential generator (LCG). Most of the existing software's today use this LCGs proposed by Lehmer in the early 50's.

Check your progress 2

E1 To find F^{-1}, we set $u = F(x)$ and solve for x to obtain

$$x = F^{-1}(u) = -\ln(1-u).$$

Therefore, to generate random variate X from exponential distribution with mean 1, first generate U from a U (0,1) and then let X = - In (1 – U). Therefore X will have exponential distribution with mean 1. Since U and 1 – U have the same U (0, 1) distribution, we can also use X = In U. This saves a subtraction.

E2 Note that $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$, therefore p_i denotes the probability mass function of a discrete random variable. The corresponding distribution function of the discrete random variable X can be written as

$$F(x) = \begin{cases} 0 & if \quad x < 1 \\ \sum_{i=1}^{m} \frac{1}{2^i} & if \quad m \le x < m+1, \end{cases}$$

where m is any positive integer. Now to generate a random deviate from the random variable X, first draw a random sample U from U (0, 1), since $0 \le U \le 1$, there exits a positive integer m such that $\sum_{i=1}^{m-1} \frac{1}{2^i} \le U < \sum_{i=1}^{m} \frac{1}{2^i}$, where $\sum_{i=1}^{0} \frac{1}{2^i} = 0$, then $X = m$,

Check your progress 3

E1 Our problem is to generate random deviate from $f(x)$ for a given $0 < \alpha < 1$. Note that we can not use the acceptance-rejection method in this case. It is easily observed if we take

$$g(x) = \begin{cases} 0 & if \quad x \le 0 \\ \frac{x^{\alpha-1}}{\Gamma(\alpha)} & if \quad 0 < x < 1 \\ \frac{e^{-x}}{\Gamma(\alpha)} & if \quad x > 1, \end{cases}$$

then $f(x) \le g(x)$ for all x. In this case

$$c = \int_{\infty}^{\infty} g(x) = \int_0^1 \frac{x^{\alpha-1}}{\Gamma(\alpha)} dx + \int_1^{\infty} \frac{e^{-x}}{\Gamma(\alpha)} dx = \frac{1}{\Gamma(\alpha)} \left[\frac{(e+a)}{ae} \right].$$

Therefore, $h(x) = \frac{1}{c}\, g(x)$ is

$$h(x) = \begin{cases} 0 & if \quad x \le 0 \\ \frac{\alpha x^{\alpha - 1}}{b} & if \quad 0 \le x \le 1 \\ \frac{\alpha e^{-x}}{b} & if \quad x > 1, \end{cases}$$

where $b = \frac{e + a}{e}$. The distribution function H (x) corresponds to the density function $h(x)$ is

$$H(x) = \int_0^x h(y)\, dy = \begin{cases} \frac{x^{\alpha}}{b} & if \quad 0 \le x \le 1 \\ 1 - \frac{\alpha e^{-x}}{b} & if \quad x > 1, \end{cases}$$

Which can be easily inverted as

$$H^{-1}(u) = \begin{cases} (bu)^{\frac{1}{\alpha}} & if \quad u \le \frac{1}{b} \\ -\ln \frac{b(1-u)}{\alpha} & if \quad u > \frac{1}{b} \end{cases}$$

Therefore, it is possible to generate random deviate from the density function $h(x)$ using the simple inversion method. Generation of a random deviate Y from the density function $h(x)$ can be performed as follows. First generate U_1 from U (0, 1), if $U_1 \le \frac{1}{b}$ we set $Y = (bU_1)^{\frac{1}{\alpha}}$, in this case $Y \le 1$. Otherwise $Y = -\ln \frac{b(1 - U_1)}{\alpha}$ and in this case $Y > 1$. Also note that

$$\frac{f(Y)}{g(Y)} = \begin{cases} e^{-Y} & if \quad 0 \le Y \le 1 \\ Y^{\alpha - 1} & if \quad Y > 1 \end{cases}$$

Now the algorithm to generate random deviate from a gamma density function with the shape parameter α, for $0 < \alpha < 1$ takes the following form:

- Step 1: Generate U_1 from U (0,1) and let $P = bU_1$. If $P > 1$, go to Step 3 otherwise proceed to Step 2.

- Step 2: Let $Y = P^{\frac{1}{\alpha}}$ and generate U_2 from U (0, 1). If $U_2 \le e^{-Y}$, return $X = Y$.
  Otherwise go back to Step 1.

- Step 3: Let $Y = -\ln \frac{(b - P)}{\alpha}$ and generate U_2 from U (0, 1). If $U_2 \le Y^{\alpha - 1}$, return $X = Y$, otherwise go back to Step 1.

UNIT 3 REGRESSION ANALYSIS

Structure **Page Nos**

3.0 INTRODUCTION

In many problems there are two or more variables that are inherently related and it may be necessary to explore the nature of their relationship. Regression analysis is a statistical technique for modeling and investigating the relationship between two or more variables. For example in a chemical process suppose that the yield of the product is related to the process operating temperature. Regression analysis can be used to build a model that expresses yield as a function of temperature. This model can be used to predict yield at a given temperature level. It can also be used for process optimization or process control purposes.

In general, suppose that there is a single dependent variable or response variable y and that is related to k independent or regressor variables say $x_1,\ldots\ldots,x_k$. The response variable y is a random variable and the regressor variables $x_1,\ldots\ldots,x_k$ are measured with negligible error. The relationship between y and $x_1,\ldots\ldots x_k$ is characterized by a mathematical model and it is known as the regression model. It is also known as the regression of y on $x_1,\ldots\ldots,x_k$. This regression model is fitted to a set of data. In many situations the experimenter knows the exact form of the functional relationship between y and $x_1, \ldots, x_k$, say $\varphi(x_1, \ldots, x_k)$, except for a set of unknown parameters. When the functional form is unknown, it has to be approximated on the basis of past experience or from the existing information. Because of its tractability, a polynomial function is popular in the literature.

In this unit we will be mainly discussing the linear regression model and when $k = 1$, that is only one regressor variables. We will be discussing in details how to estimate the regression line and how it can be used for prediction purposes from a given set of data. We will also discuss briefly how we can estimate the function ϕ, if it is not linear.

3.1 OBJECTIVES

After reading this unit, you should be able to

- Decide how two variables are related.
- Measure the strength of the linear relationship between two variables.
- Calculate a regression line that allows to predict the value of one of the variable if

the value of the other variable is known.

- Analyze the data by the method of least squares to determine the estimated regression line to be used for prediction.
- Apply the least squares methods to fit different curves and use it for prediction purposes.

3.2 SIMPLE LINEAR REGRESSION

We wish to determine the relationship between a single regressor variable *x* and a response variable *y(note: The linear regression with one independent variable is referred to as simple linear regression)*. We will refer to y as the dependent variable or response and x as the independent variable or regressor. The regressor variable *x* is assumed to be a continuous variable controlled by the experimenter. You know that it is often easy to understand data through a graph. So, let us plot the data on *Scatter diagram* (a set of points in a 2-D graph where horizontal axis is regressor and vertical axis is response) Suppose that the true relationship between *y* and *x* is straight line. Therefore, each observation y can be described by the following mathematical relation (model)

$$y = \beta_0 + \beta_1 x + \in \qquad (1)$$

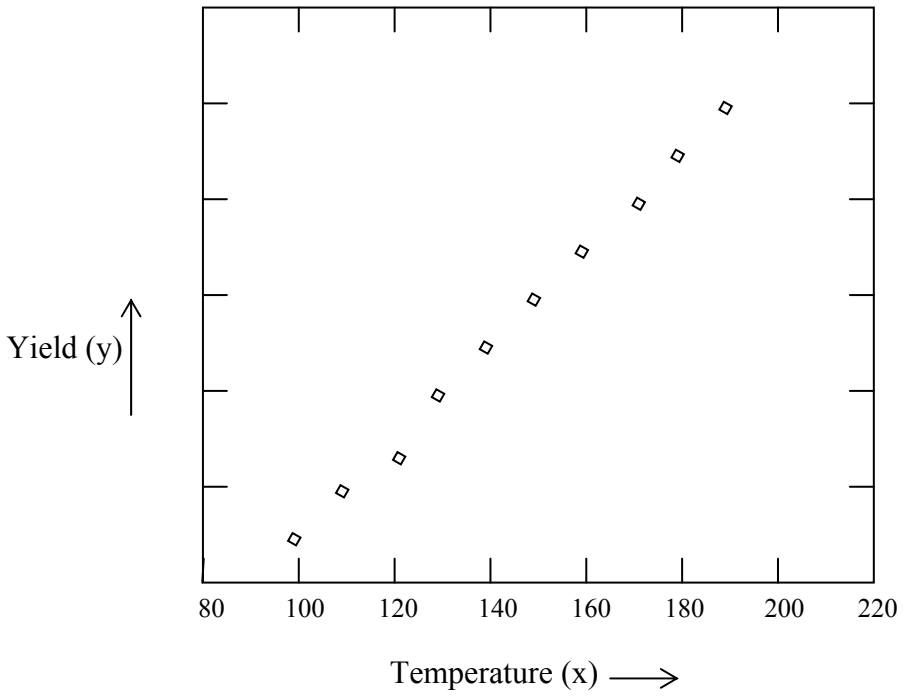


Figure 1: Scatter diagram of yield versus temperature

where $\in$ is a random variable with mean 0 and variance σ^2. The $\in$ is known as the error component and it is assumed to be small. If the error $\in$ was absent then it was a perfect relation between the variables *y* and *x* which may not be very practical. Let us look at the following example.

Example 1: A chemical engineer is investigating the effect of process operating temperature on product yield. The study results in the following data.

Temperature °C (x)	100	110	120	130	140	150	160	170	180	190
Yield,%(y)	45	51	54	61	66	70	74	78	85	89

The scatter diagram between the temperature and the yield is presented in the Figure 1 above. From the Figure 1 it is clear that there is a linear relationship between yield and temperature but clearly it is not perfect. For example we can not write the relationship between y and x as follows

$$y = \beta_0 + \beta_1 x$$

Clearly the presence of the error $\in$ is needed. Moreover the error $\in$ is a random variable because it is not fixed and it varies from one temperature to another. It may also vary when two observations are taken at the same temperature. If there was a perfect linear relationship between y and x we would have required just two points to find the relationship. Since the relationship is not perfectly linear it is usually required much more than two data points to find their relationship. Our main objective is to find the relationship between them from the existing information (data points). Since it is assumed that the relationship between x and y is linear therefore the relationship can be expressed by the equation (1) and finding the relationship basically boils down finding the unknown constants β_0 and β_1 from the observations.

Let us discuss this concept of linear regression by one more illustration/collection of data described in the table 1 given below. This table encloses the data of 25 samples of cement, for each sample we have a pair of observation (x,y) where x is percentage of SO_3, a chemical and y is the setting time in minutes. These two components are strongly related; it is the percentage of SO_3 which influences the setting time of any cement sample, the recorded observations are given in table 1 below.

Table 1: Data on SO_3 and Setting Time

S.No. i	Percentage of SO_3 x	Setting Time Y (in minutes)
1	1.84	190
2	1.91	192
3	1.90	210
4	1.66	194
5	1.48	170
6	1.26	160
7	1.21	143
8	1.32	164
9	2.11	200
10	0.94	136
11	2.25	206
12	0.96	138
13	1.71	185
14	2.35	210
15	1.64	178
16	1.19	170
17	1.56	160
18	1.53	160

19	0.96	140
20	1.7	168
21	1.68	152
22	1.28	160
23	1.35	116
24	1.49	145
25	1.78	170
Total	39.04	4217
Sum of Squares	64.446	726539

From the table 1, you see that setting time y increases as percentage of SO_3 increases. Whenever you find this type of increasing (or decreasing) trend in a table , same will be reflected in the scatter diagram ,and it indicates that there is a linear relationship between x and y. By drawing the scatter diagram you can observe that the relationship is not perfect in the sense that a straight line cannot be drawn through all the points in the scatter diagram.

Nevertheless, we may approximate it with some linear equation. What formula shall we use? Suppose, we use the formula y = 90 + 50x to predict y based on x. To examine how good this formula is, we need to compare the actual values of y with the corresponding predicted values. When x = 0.96, the predicted y is equal to 138(= 90 + 50 *x* 0.96). Let (x_i, y_i) denote the values of (x, y) for the i^{th} sample. From Table-1, notice that $x_{12}=x_{19}=0.96$, whereas $y_{12} = 138$ and $y_{19} = 140$. Let $\hat{y} = 90 + 50x_i$. That is, $\hat{y}_i$ **is the predicted value of y** (then using y = 90 + 50x for the i^{th} sample. Since, $x_{12}=x_{19}=0.96$, both $\hat{y}_{12}$ and $\hat{y}_{19}$ are equal to 138. Thus the difference $\hat{e}_i = y_i - \hat{y}_i$, **the error in prediction**, also called residual is observed to be $\hat{e}_{12} = 0$ and $\hat{e}_{19} = 2$. The formula we have considered above, y = 90 + 50x, is called a **simple linear regression equation** , we will study these terms in detail in our successive sections.

3.2.1 Least squares estimation

Suppose that we have *n* pairs of observations, say (x_1 , y_1),…….(x_n , y_n). It is assumed that the observed y_i and x_i satisfy a linear relation as given in the model (1). These data can be used to estimate the unknown parameters β_0 and β_1 . The method we are going to use is known as the method of least squares, that is, we will estimate β_0 and β_1 so that the sum of squares of the deviations from the observations to the regression line is minimum. We will try to explain it first using a graphical method in Figure 2. For illustrative purposes we are just taking 5 data points (*x*, *y*) = (0.5, 57), (0.75, 64), (1.00, 59), (1.25, 68), (1.50,74). The estimated regression line can be obtained as follows. For any line we have calculated the sum of the differences (vertical distances) squares between the *y* value and the value, which is obtained using that particular line. Now the estimated regression line is that line for which the sum of these differences squares is minimum.

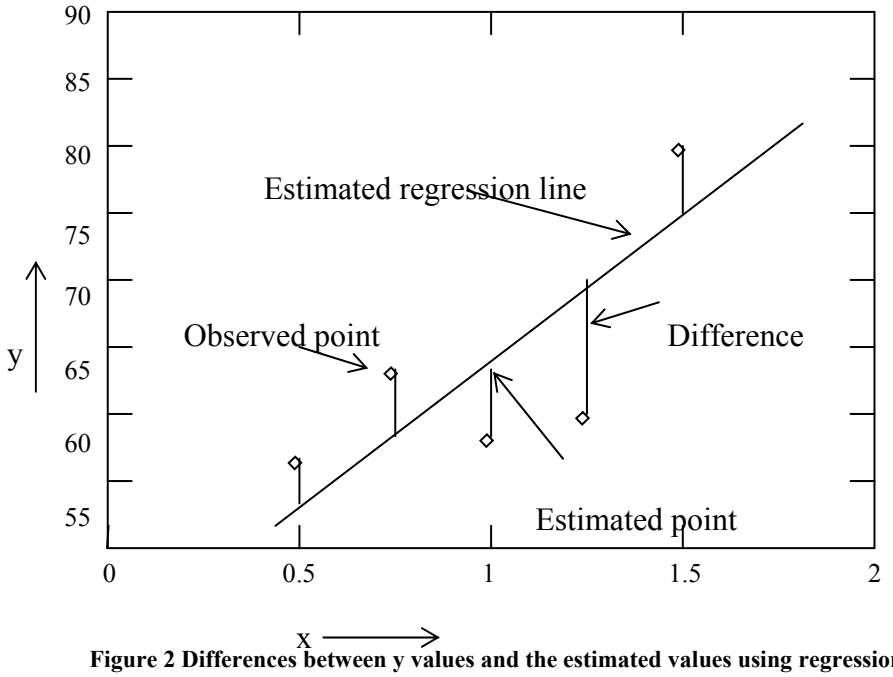


Figure 2 Differences between y values and the estimated values using regression line

Matematically the sum of squares of the deviations of the observations from the regression line is given by

$$L = \sum_{i=1}^{n} \in_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimators of β_0 and β_1, are $\hat{\beta}_0$ and $\hat{\beta}_1$ which can be obtained by solving the following two linear equations.

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

Simplifying these two equations yields

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (2)$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \qquad (3)$$

Solving (2) and (3) $\hat{\beta}_0$ and $\hat{\beta}_1$ can be obtained as follows:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (4)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\left(\sum_{i=1}^{n} yi\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} xi\right)^2} \qquad (5)$$

where $\bar{y} = \sum_{i=1}^{n} y_i$ and $\bar{x} = \sum_{i=1}^{n} x_i$. Therefore, the fitted simple linear regression line between these *n* points is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (6)$$

Note : Linear Correlation and Regression

linear correlation and regression are very similar. One uses the correlation coefficient to determine whether two variables are linearly related or not. The correlation coefficient measures the strength of the linear relationship. Regression on the other hand is used when we want to answer question about the relationship between two variables.

Some Properties for Linear Correlation and Regression

(1) The line of regression of y on x always passes through($\bar{x}$, $\bar{y}$) where $\bar{x}$, and $\bar{y}$ are the mean of x and y values.

(2) There are always two line of regression one of y on x and other of x on y. i.e., $y = a_1 + b_{yx}\, x$ or $x = a_2 + b_{xy}\, y$

where byx = Regression coeff of y on x $= r\frac{\sigma_y}{\sigma_x}$

bxy = Regression coeff of y on x $= r\frac{\sigma_x}{\sigma_y}$

Correlation can be obtained by the following formula also,

$r = \sqrt{b_{xy} * b_{yx}}$ $(-1 \le r \le 1)$

Angle between lines of regression is given by,

$$\theta = \tan^{-1}\left\{\frac{r^2 - 1}{r}\left(\frac{\sigma_x * \sigma_y}{\sigma^2{}_x + \sigma^2{}_y}\right)\right\}$$

Where r = correlation coeff between x and y

σ_x = standard deviation of variable x

σ_y = standard deviation of variable y

So, now, Regression equation of y on x can be rewritten as

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

And Regression equation of x on y as,

$$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Example 1 (contd.) Now we will compute the estimates of β_0 and β_1 for the data points given in Example 1. In this case it is observed

$$n = 10, \qquad \sum_{i=1}^{10} x_i = 1450, \qquad \sum_{i=1}^{10} yi = 673, \quad \bar{x} = 145, \qquad \bar{y} = 67.3$$

$$\sum_{i=1}^{10} x_i^2 = 218,500, \quad \sum_{i=1}^{10} y_i^2 = 47,225, \quad \sum_{i=1}^{10} x_i y_i = 101,570$$

Therefore,

$$\hat{\beta}_1 = \frac{101,570 - 10 \times 1450 \times 673}{218,500 - 10 \times 1450^2} = 0.483,$$

and

$$\hat{\beta}_0 = 673 - 0.483 \text{ X } 145 = -2.739.$$

The fitted simple linear regression line through those 10 points is

$$\hat{y} = -2.739 + 0.483x \qquad (7)$$

The best fitted line along with the data points are plotted in the Figure 3. Note that the best fitted line can be used effectively for prediction purposes also. For example suppose we want to know the expected yield when the temperature is 170°C, for which the data is not available. We can use the (7) for this purpose as follows.

$$\hat{y} = -2.739 + 0.483 \times 170 = 79.371 .$$

Therefore, the best fitted line shows that the expected yield at 170°C is 79.371.

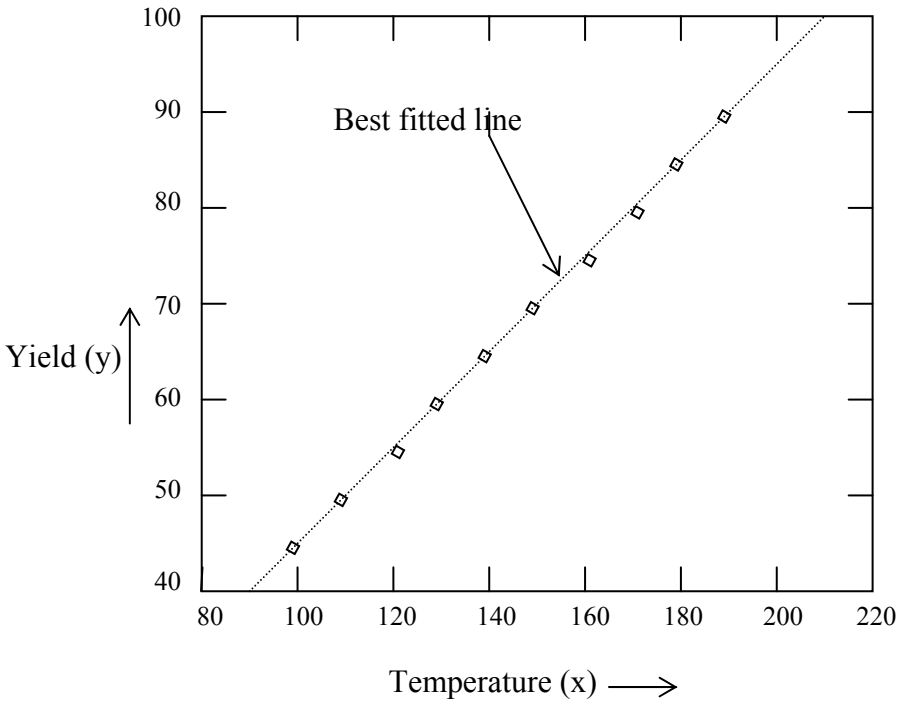


Figure 3: Data points and the best fitted regression line passing through these points

Soon in this section only, we will discuss the technique consisted of few steps; which can be used to fit a line in best way, such that the error is minimum. In crux we will study the technique to determine the best equation, that can fit a line in the data such that the error is minimum. But before that lets see one more example.

Example 2: A survey was conducted to relate the time required to deliver a proper presentation on a topic , to the performance of the student with the scores he/she receives. The following Table shows the matched data:

Table 2

Hours (x)	Score (y)
0.5	57
0.75	64
1.00	59
1.25	68
1.50	74
1.75	76
2.00	79
2.25	83
2.50	85
2.75	86
3.00	88
3.25	89
3.50	90
3.75	94
4.00	96

(1) Find the regression equation that will predict a student's score if we know how many hours the student studied.

(2) If a student had studied 0.85 hours, what is the student's predicted score?

Solution. We will arrange the data in the form of a chart to enable us to perform the computations easily.

Table 3

x	y	x^2	xy
0.5	57	0.25	28.5
0.75	64	0.56	48.0
1.00	59	1.00	59.0
1.25	68	1.56	85.0
1.50	74	2.25	111.0
1.75	76	3.06	133.0
2.00	79	4.00	158.0
2.25	83	5.06	186.75
2.50	85	6.25	212.5
2.75	86	7.56	236.5
3.00	88	9.00	246.0
3.25	89	10.56	289.25
3.50	90	12.25	315.0
3.75	94	14.06	352.50
4.00	96	16.00	384.0
33.75	1188	93.44	2863

In this case n = 15, therefore

$$\hat{\beta}_1 \frac{15\times 2863 - 33.75\times 1188}{15\times 93.44 - 33.75^2} = 10.857\,,\ \hat{\beta}_0 = \frac{1}{15}[1188 - 10.857\times 33.75] = 54.772$$

Therefore, the prediction line becomes:

$$\hat{y} = 54.772 + 10.857x$$

Now the predicted score when $x = 0.85$, is

$$\hat{y} = 54.772 + 10.857 \times 0.85 = 64.00$$

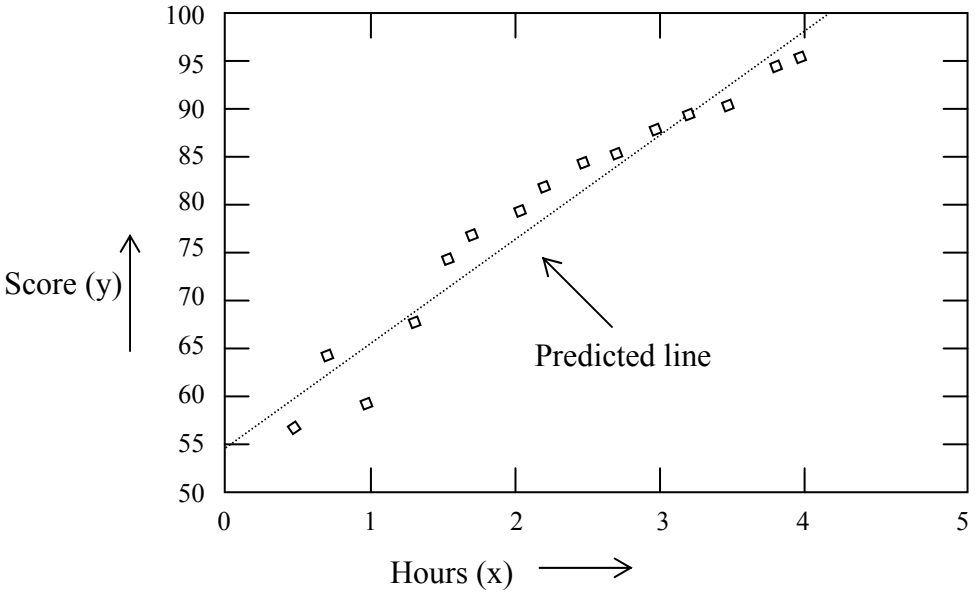


Figure 4: Hours studied and the corresponding score with the best fitted regression line passing through these points

Thus the predicted score of the student who had studied 0.85 hours is approximately 64.00.

We have plotted the individual scores and the hours studied with the best fitted prediction line in the Figure 4. It shows the hours studied by the student and the corresponding score follows a linear pattern and the predicted line can be used quite effectively to predict the score of a student if we know how many hours the student had studied.

Now, it's the time to discuss the technique for determining the best equation, i.e., the equation which fits the line in a way that the overall error is minimized.

From above illustrations and examples you might have noticed that different equations give us different residuals. What will be the best equation? Obviously, the choice will be that equation for which $\hat{e}_i s$ are small.

This means that whatever straight line we use, it is not possible to make all $\hat{e}_i s$ zero , where $\hat{e}_i = y_i - \hat{y}_i$ (the difference). However, we would expect that the errors are positive in some cases and negative in the other cases so that, on the whole, their sum is close to zero. So, our job is to find out the best values of β_0 and β_1 in the formula $y = \beta_0 + \beta_1 x + e$ $(s.t. \quad e \approx 0)$. Let us see how we do this.

Now our aim is to find the values β_0 and β_1 so that the error $\hat{e}_i s$ are minimum. For that we state here four steps to be done.

1) Calculate a sum S_{xx} defined by

$$S_{xx} = \sum_{i=1}^{n} x^2_i - n\bar{x}^2 \qquad (8)$$

where x_i's are given value of the data and $\bar{x} = \frac{\Sigma x_i}{n}$ is the mean of the observed values and n is the sample size.

The sum S_{xx} is called the corrected sum of squares.

2) Calculate a sum S_{xy} defined by

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \qquad (9)$$

where x_i's and y_i's are the x-values and y-values given by the data and $\bar{x}$ and $\bar{y}$ are their means.

3) Calculate $\frac{S_{xy}}{S_{xx}} = \beta_1$ say. That is

$$\beta_1 = \frac{S_{xy}}{S_{xx}} \qquad (10)$$

4) Find $\bar{y} - \beta_1 \bar{x} = \beta_0$, say.

Let us now compute these values of the data in Table 1: Data on SO_3 and Setting Time, we get

$\bar{x}$ = 1.5616, $\bar{y}$ = 168.68, S_{xx} = 3.4811, and S_{xy} = 191.2328.

Substituting these values in (10) and (11), we get

$$\beta_1 = \frac{S_{xy}}{S_{xx}} = 54.943 \text{ and } \beta_0 = 168.68 - 54.943 \text{ x } 1.5616 = 82.88 \qquad (11)$$

Therefore, the best linear prediction formula is given by

y = 82.88 + 54.943x.

After drawing this line on the scatter diagram, you can find that this straight lines is close to more points, and hence it is the best linear prediction.

Example 3: A hosiery mill wants to estimate how its monthly costs are related to its monthly output rate. For that the firm collects a data regarding its costs and output for a sample of nine months as given in the following table.

Table 4

Output (tons)	**Production cost (thousands of dollars)**
1	2
2	3
4	4
8	7
6	6
5	5
8	8
9	8
7	6

(a) Find the scatter diagram for the data given above.
(b) Find the regression equation when the monthly output is the dependent variable (x) and monthly cost is the independent variable (y).
(c) Use this regression line to predict the firm's monthly cost if they decide to produce 4 tons per month.
(d) Use this regression line to predict the firm's monthly cost if they decide to produce 9 tons per month.

Solution

a) Suppose that x_i denote the output for the ith month and y_i denote the cost for the i^{th} month. Then we can plot the graph for the pair (x_i, y_i) of the values given in Table . Then we get the scatter diagram as shown in Figure below.

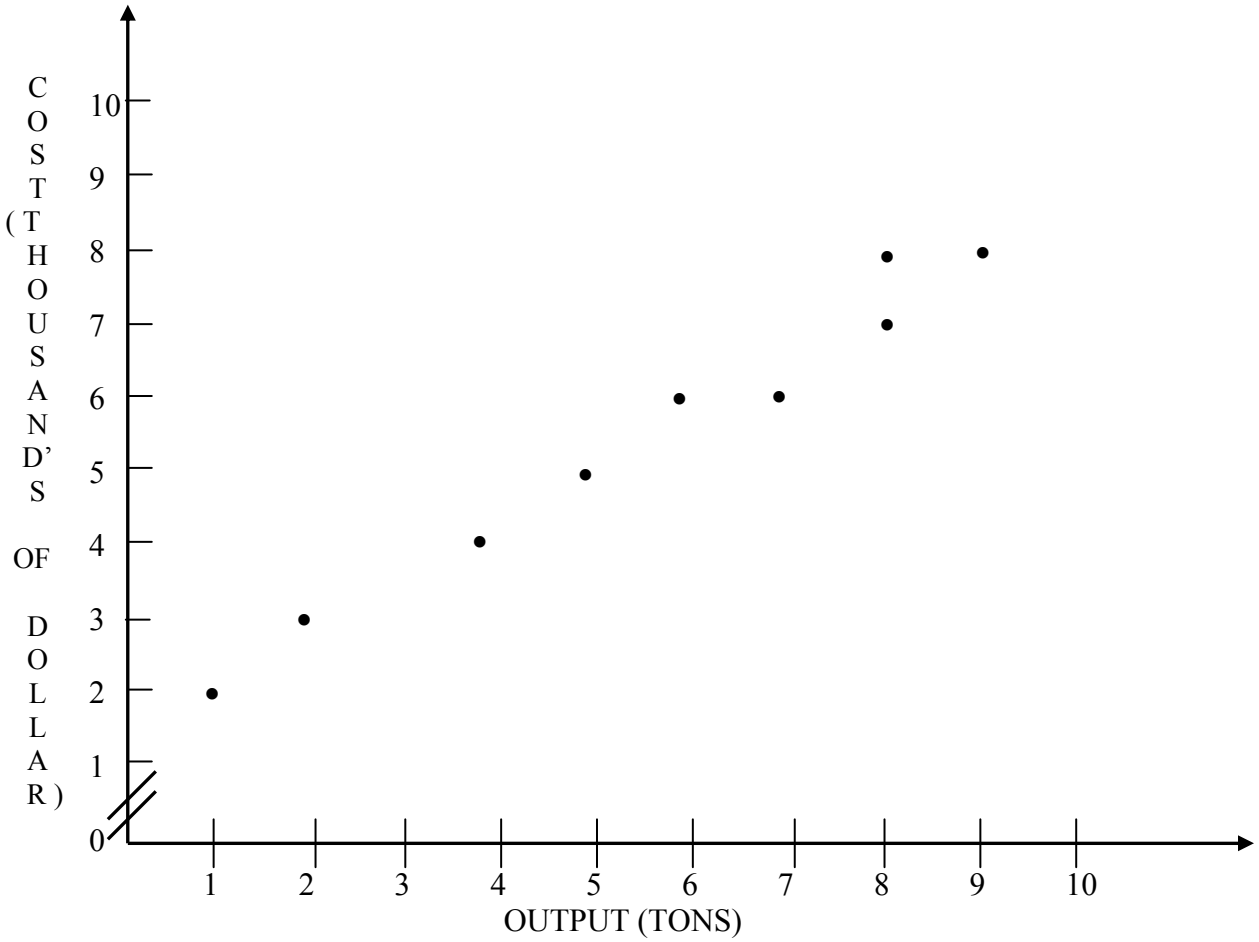


Figure 5: Scatter Diagram

b) Now to find the least square regression line, we first calculate the sums S_{xx} and S_{xy} from Eqn.(8) and (9).

$$S_{xx} = \sum_{i=1}^{n} x^2{}_i - n\bar{x}$$

Note that from Table(4) we get that

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{50}{9}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{49}{9}$$

$$\sum x^2{}_i = 340$$

$$\sum y^2{}_i = 303$$

and $\sum x_i y_i = 319$

Therefore, we get that

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_1{}^2 - n\bar{x}^2}$$

$$= \frac{9\ x\ 319 - 50\ x\ 49}{9\ x\ 340 - 50^2}$$

$$= \frac{421}{560} = 0.752$$

Correspondingly, we get

$$\beta_0 = \frac{49}{9} - (0.752)\ x\ \frac{50}{9}$$

= 1.266

Therefore, the best linear regression line is

y = 1.266 + (0.752)x

c) If the firms decides to produce 4 tons per month, then one can predict that its cost would be

1.266 + (0.752) x 4 = 4.274

Since the costs are measured in thousands of dollars, this means that the total costs would be expected to be $4,274.

d) If the firms decides to produce 9 tons per month, then one can predict that its cost would be 1.266 + (0.752) x 9=8.034

Since the costs are measured in thousands of dollars, this means that the total costs would be expected to be $8034.

Check your progress 1

E 1: In partially destroyed laboratory record of an analysis of correlation data, the following results only are legible.

Variance of x = 9

Regression equations : 8x – 10y + 66 = 0

40x – 18y – 214 = 0

what were (1) the mean values of x and y,
(2) the correlation coeff between x and y
(3) the standard deviation of y

E2 Since humidity influences evaporation, the solvent balance of water reducible paints during sprayout is affected by humidity. A controlled study is conducted to examine the relationship between humidity (X) and the extent of evaporation (Y) is given below in table 5. Knowledge of this relationship will be useful in that it will allow the painter to

adjust his or her spray gun setting to account for humidity. Estimate the simple linear regression line and predict the extent of solvent evaporation (i.e loss of solvent ,by weight)when the relative humidity is 50%

Table 5

Observation	(x) Relative humidity, (%)	(y) Solvent Evaporation, (%) wt
1	35.3	11.0
2	29.7	11.1
3	30.8	12.5
4	58.8	8.4
5	61.4	9.3
6	71.3	8.7
7	74.4	6.4
8	76.7	8.5
9	70.7	7.8
10	57.5	9.1
11	46.4	8.2
12	28.9	12.2
13	28.1	11.9
14	39.1	9.6
15	46.8	10.9
16	48.5	9.6
17	59.3	10.1
18	70.0	8.1
19	70.0	6.8
20	74.4	8.9
21	72.1	7.7
22	58.1	8.5
23	44.6	8.9
24	33.4	10.4
25	28.6	11.1

3.2.2 Goodness of Fit

We have seen in the previous subsection that the regression line provides estimates of the dependent variable for a given value of the independent variable. The regression line is called the best fitted line in the sense of minimizing the sum of squared errors. The best fitted line shows the relationship between the independent (x) and dependent (y) variables better than any other line. Naturally the question arises “How good is our best fitted line?”. We want a measure of this goodness of fit. More precisely we want to have a numerical value which measures this goodness of fit.

For developing a measure of goodness of fit, we first examine the variation in *y*. Let us first try the variation in the response *y*. Since *y* depends on *x*, if we change *x,* then *y* also changes. In other words, a part of variation in *y*’s is accounted by the variation in *x*’s.

Actually, we can mathematically show that the total variation in y's can be split up as follows:

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{S^2_{xy}}{S_{xx}} + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad (12)$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2; S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \hat{y}_i)$$

Now if we divide (12) by S_{yy} on both sides, we get

$$1 = \frac{S^2_{xy}}{S_{xx}S_{yy}} + \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{S_{yy}}$$

Since the quantities on the right hand side are both non-negative, none of them can exceed one. Also if one of them is closer to zero the other one has to be closer to one. Thus if we denote

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

then

$$R^2 = \frac{S^2_{xy}}{S_{xx}S_{yy}}$$

Since R^2 must be between 0 and 1, R must be between – 1 and 1. It is clear that if $R^2 = 1$, then

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{S_{yy}} = 0 \text{ or } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0 \quad \text{or} \quad y_i = \hat{y}_i \quad \text{for all } i.$$

Again when R^2 is close to 1, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is close to zero. When R is negative, it means that y decreases as x increase and when R is positive y increases when x increases. Thus R gives a measure of strength of the relationship between the variables x and y.

Now let us compute the value of R for Example 1. For calculating the numerical value of R, the following formula can be used;

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}}$$

Therefore, for Example 1, the value of R becomes;

$$\text{R} = \frac{101,570 - 10 \times 145 \times 67.3}{\sqrt{218,500 - 10 \times 145^2}\sqrt{47225 - 10 \times 67.3^2}} = \frac{3985}{\sqrt{8250}\sqrt{1932.1}} = 0.9981$$

and $R^2 = 0.9963$.

Therefore, it is clear from the value of R or from R^2 that both of them are very close to one. From the figure also it is clear that the predicted line fits the data very well.

Moreover R is positive means, there is a positive relation between the temperature and yield. As the temperature increases the yield also increases.

Now the natural question is how large this R or R^2 will be to say that the fit is very good. There is a formal statistical test based on F-distribution which can be used to test whether R^2 is significantly large or not. We are not going into that details. But as a thumb rule we can say that if R^2 is greater that 0.9, the fit is very good, if it is between 0.6 to 0.8, the fit is moderate and if it is less than 0.5 it is not good.

Check your progress 2

E1) For the data given in the table below compute R and R^2

Table 6: $\hat{y}_i$ and $\hat{e}_i$ For Some Selected i

Sample No. (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
$\hat{y}_i$	138				
$\hat{e}_i$	0				

Note: $\hat{y}_i = 90 + 50x$ and $\hat{e}_i = y_i - \hat{y}_i$

3.2.3 Residual Analysis

Fitting a regression model to a set of data requires several assumptions. For example estimation of the model parameters requires the assumptions that the errors are uncorrelated random variables with mean zero and constant variance. If these assumptions are not satisfied, then using the simple least squares method may not produce the efficient regression line. In fitting a linear model, it is also assumed that the order of the model is correct, that is if we fit a first order polynomial, then we are assuming that phenomenon actually behave in a first order manner. Therefore, for a practitioner it is important to verify these assumptions and the adequacy of the model.

We define the residuals as $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$, where y_i is an observation and $\hat{y}_i$ is the corresponding estimated value from the best fitting regression line.

Analysis of the residuals is frequently helpful in checking the assumption that errors are independent and identically distributed with mean zero and finite variance and in determining `whether the additional terms in the model would be required not. It is advisable to plot the residuals

a) in time sequence (if known),
b) against the $\hat{y}_i$ or
c) against the independent variable x. These graphs will usually look like one of the four general patterns as shown in the Figures 6 – 9.

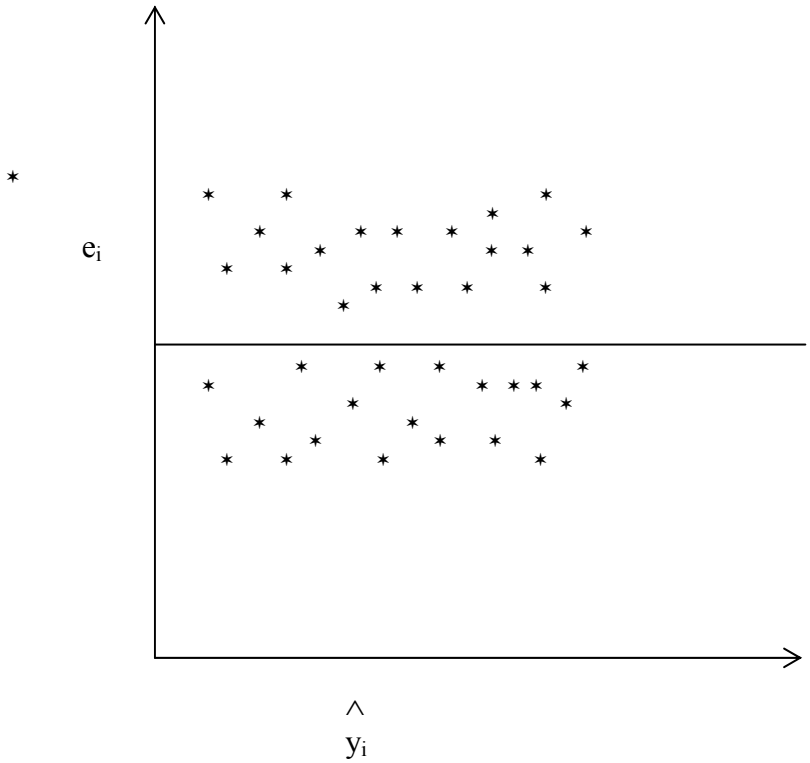


Figure 6 Pattern of the residual plot; satisfactory.

The Figure 6 indicates that the residuals are behaving in satisfactory manner and the model assumptions can be assumed to be correct. The Figures 7 – 9 given below indicate unsatisfactory behaviour of the residuals. The Figure 7 clearly indicates that the variances are gradually increasing. Similarly the Figure 8 indicates that the variances are not constant. If the residuals plot is like the Figure 9, then it seem the model order is not correct, that means, the first order model is not the correct assumption. We should look for higher order models.

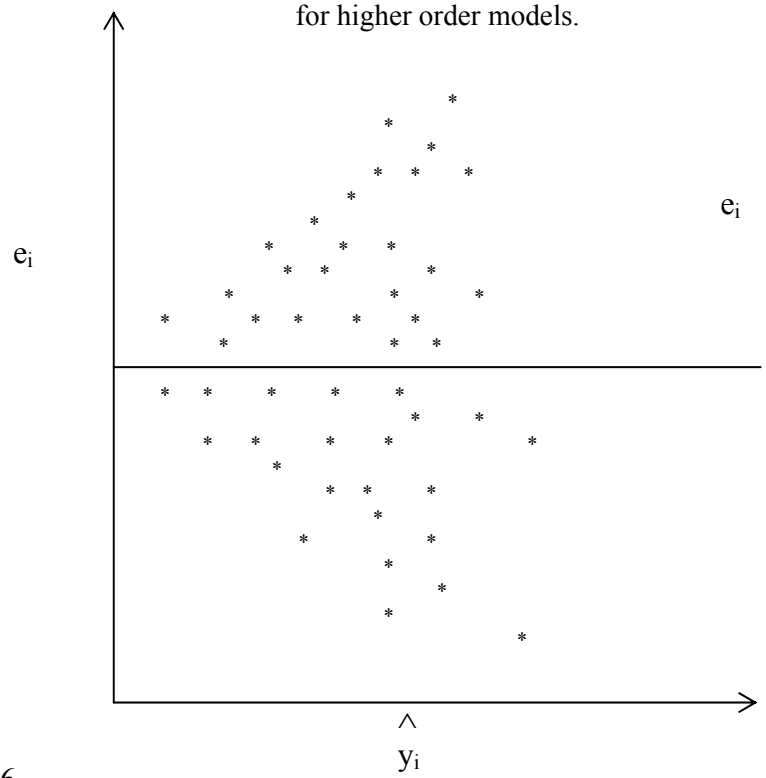


Figure 7 Pattern of the residual plot; indicates the variance is gradually increasing this case

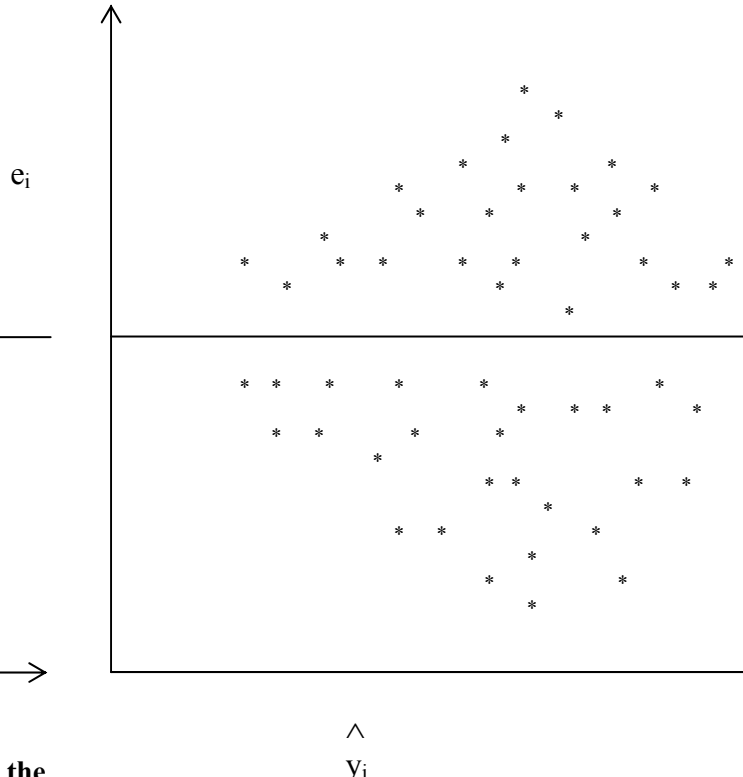


Figure 8 Pattern of the residual plot; indicates the vari not constant.

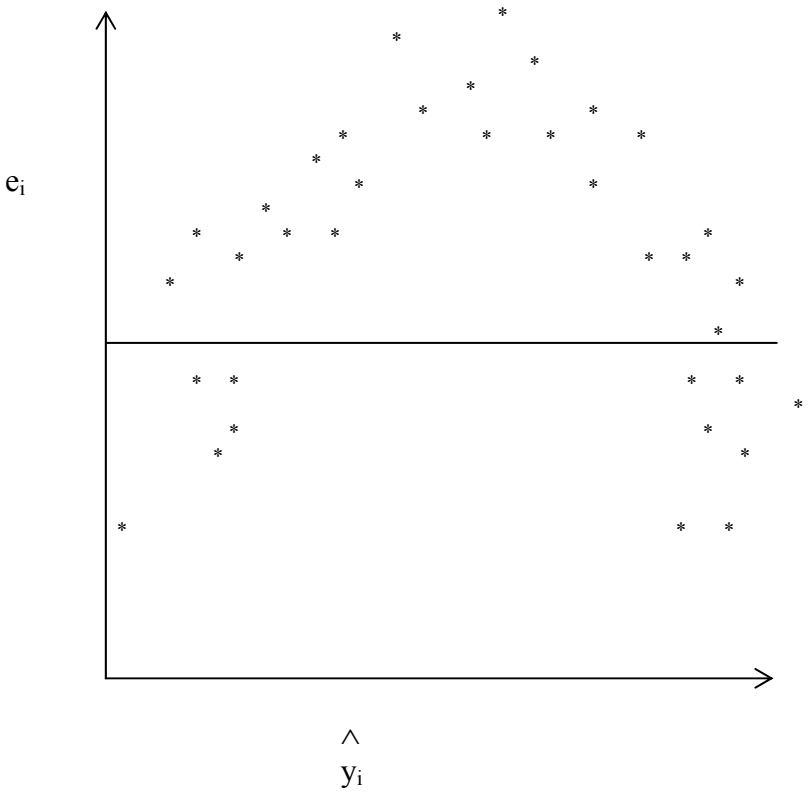


Figure 9 Patent of the residual plot; indicates the model order is not correct.

Example 4: Now we provide the residual plots of the data given in Example 1. We have plotted $\hat{y}_i$ vs. e_i. It is provided in the Figure 10. From the Figure 10, it is quite clear that the residuals plot is quite satisfactory and apparently all the model assumptions are satisfied in Figure 10 here.

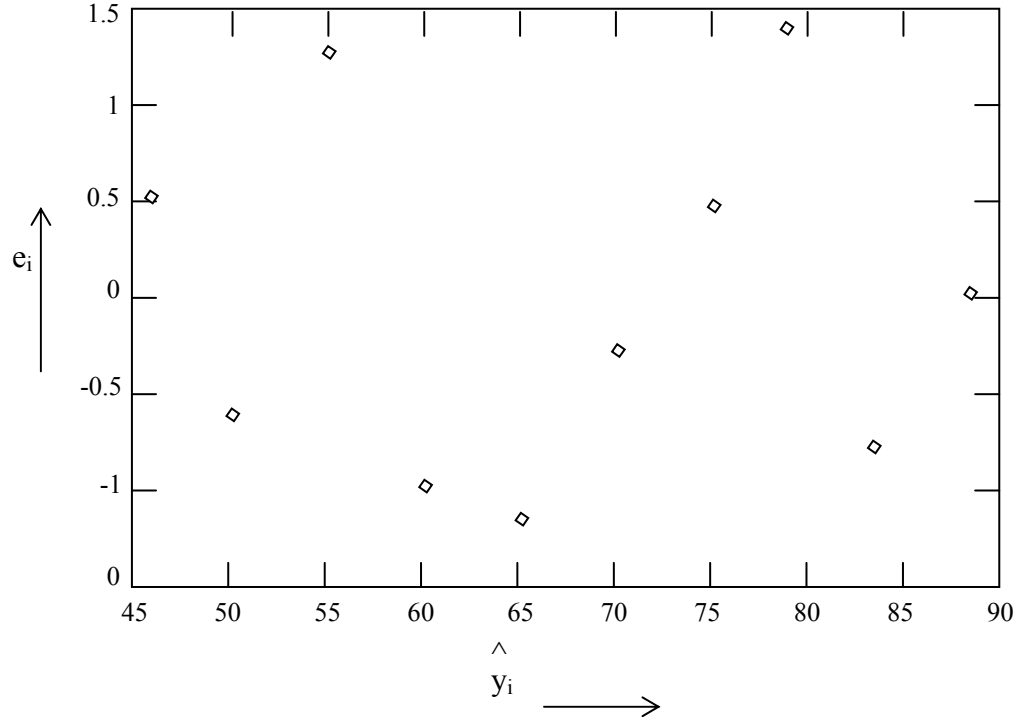


Figure10 Pattern of the residual plot; satisfactory.

Check your progress 3

E1 What is the utility of residual plots? what is the disadvantage of residual plots?

3.3 NON-LINEAR REGRESSION

Linear regression is a widely used method for analyzing data described by models which are linear in parameters. However, in many practical situations, people come across with data where the relationship between the independent variable and the dependent variable is no more linear. In that case definitely one should not try to use a linear regression model to represent the relationship between the independent and dependent variable. Let us consider the following example.

Example 5 Data on the amount of protein generated by a certain experiment were counted and reported over time. They are presenting below in Table 7:

Table 7

Time (min)	Protein (gm)	Time (min)	Protein (gm)	Time (min)	Protein (gm)	Time (min)	Protein (gm)
0	0.844	80	0.818	160	0.580	240	0.457
10	0.908	90	0.784	170	0.558	250	0.448
20	0.932	100	0.751	180	0.538	260	0.438
30	0.936	110	0.718	190	0.522	270	0.431
40	0.925	120	0.685	200	0.506	280	0.424
50	0.908	130	0.685	210	0.490	290	0.420
60	0.881	140	0.628	220	0.478	300	0.414
70	0.850	150	0.603	230	0.467	310	0.411

We present the Time vs. Protein generated in the Figure 11.

From Figure 11 it is clear that the relationship between the time and protein generated is not linear. Therefore, they can not be explained by a linear equation. In a situation like this we may often go for a non-linear model to explain the relationship between the independent and dependent variables and they are called the non-linear regression model.

A non-linear regression model can be formally written as

$$y = f(x,\theta) + \in , \tag{13}$$

where $f(x,\theta)$ is a known response function of k-dimensional vector of explanatory variable x and p-dimensional vector of unknown parameter θ. Here also $\in$ represents the error component and it is assumed that it has mean zero and finite variance. Therefore, it is clear that the non-linear regression model is a generalization of the linear regression model. In case of linear regression model $f(x, \theta)$ is a linear function, but there it can be any non-linear function also. Similar to the linear regression model, here also our problem is same, that is, if we observe a set of n, $\{(x_1, y_1),\ldots\ldots, (x_n, y_n)\}$ how to estimate the unknown parameters θ, when we know the functional form of $f(x, \theta)$.

3.3.1 LEAST SQUARES ESTIMATION

Similar to the linear regression method here also to estimate the unknown parameters, we adopt the same method. We find the estimate of θ by minimizing the residual sums of squares, that is minimize

$$Q(\theta) = \sum_{i=1}^{n} [y_i - f(x_i, \theta)]^2, \qquad (14)$$

with respect to the unknown parameters. The idea is same as before, that is we try to find that particular value of θ for which the sum of squares of the distance between the points y_i and $f(x_i, \theta)$ is minimum. Unfortunately in this case the minimum can not be performed as easily as before. We need to adopt some numerical technique to minimize the function $Q(\theta)$. This minimization can be performed iteratively and one technique that can be used to accomplish this is the Gauss-Newton method.

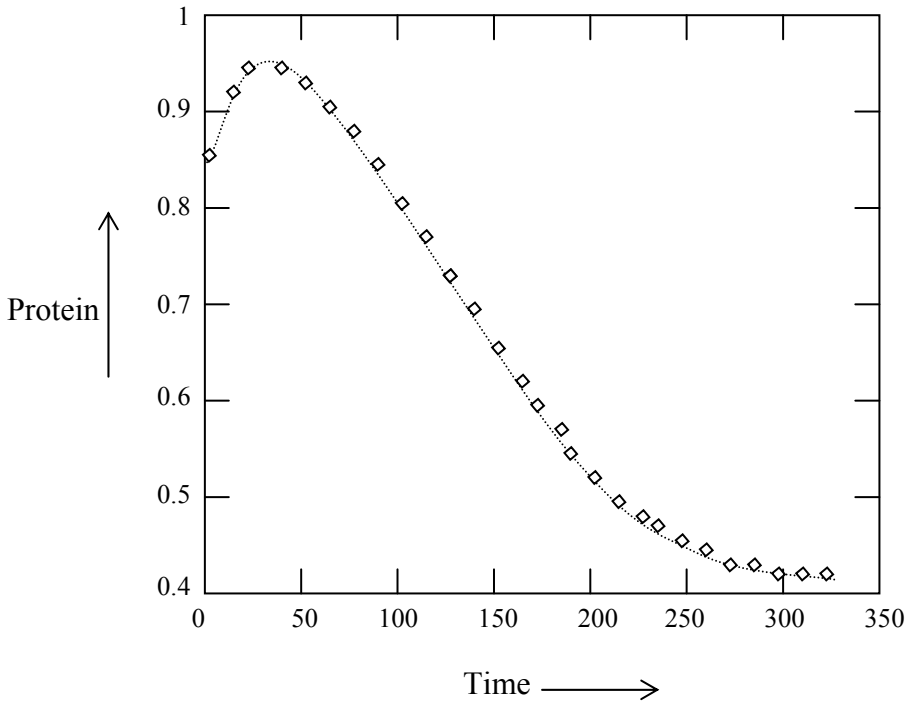


Figure 11 Time vs. Protein generated in an experiment.

You have already learned about the Gauss-Newton method in details before, we just give a brief description for your convenience. We use the following notations below:

$$\theta = (\theta_1, ..., \theta_p), \qquad \theta^{(k)} = (\theta_1^{(k)}, ..., \theta_p^{(k)}) \qquad (15)$$

Expand the function $f(x, \theta)$ using a Taylor series expansion about the starting point $\theta^{(0)}$ and using the only the first order expansion, we get:

$$f(x_i, \theta) \approx f(x_i, \theta^{(0)}) + v_{i1}(\theta_1 - \theta_1^{(0)}) + ... + v_{ip}(\theta_p - \theta_p^{(0)})$$

where

$$v_{ij} = \left. \frac{\partial f(x_i, \theta)}{\partial \theta_j} \right|_{\theta = \theta^{(0)}} \qquad \text{for} \qquad j = 1, ..., p.$$

Let $\eta(\theta) = (f(x_1,\theta),.....,f(x_n,\theta))'$ and $y = y = (y_1,.....y_n)'$ then in the matrix notation we can write (15)

$$\eta(\theta) \approx \eta(\theta^{(0)}) + V^{(0)}(\theta - \theta^{(0)}),$$

where $V^{(0)}$ is the $\eta \times p$ derivative matrix with elements v_{ij}. Therefore, to compute the first estimates beyond the starting value is to compute

$$b_0 = [V^{(0)'}V^{(0)}]^{-1}[y - \eta(\theta^{(0)})]$$

and then solve for the new estimate $\theta^{(1)}$ as

$$\theta^{(1)} = b_0 + \theta^{(0)}.$$

This procedure is repeated then with $\theta^{(0)}$ is replaced by $\theta^{(1)}$ and $V^{(0)}$ by $V^{(1)}$ and this produces a new set of estimates. This iterative procedure continues until convergence is achieved. Naturally these computations can not performed by hands, we need calculators or computers to perform these computations.

Example 5 (Contd). In this case it is observed (theoretically) that the following model (16) can be used to explain the relationship the time and yield generated y_i where

$$y_t = \alpha_0 + \alpha_1 e^{\beta_1 t} + \alpha_2 e^{\beta_2 t} + \in_t . \qquad (12)$$

Note that as we have mentioned for the general non-linear regression model, in this case also the form of the non-linear function namely $\alpha_0 + \alpha_1 e^{\beta_1 t} + e^{\beta_2 t}$ is known, but the parameters of the model, that is, $\theta = (\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$ is unknown. Given the data as provided in Example 5, we want to estimate the unknown parameters.

We use the following initial guess $\alpha_0 = 0.5, \alpha_1 = 1.5, \alpha_2 = -1.0, \beta_1 = -0.01, \beta_2 = -0.02,$ and finally using the Gauss-Newton algorithm we obtain the estimates of the parameters as follows:

$\hat{\alpha}_0 = 0.375,$ $\hat{\alpha}_1 = 1.936$ $\hat{\alpha}_2 = 1.465,$ $\hat{\beta}_0 = -0.013$ $\hat{\beta}_1 = -0.022$

We have plotted the points and also the best fitted regression line, namely

$$\hat{y} = 0.375 + 1.936e^{-0.013t} - 1.465e^{-0.013t} \qquad (17)$$

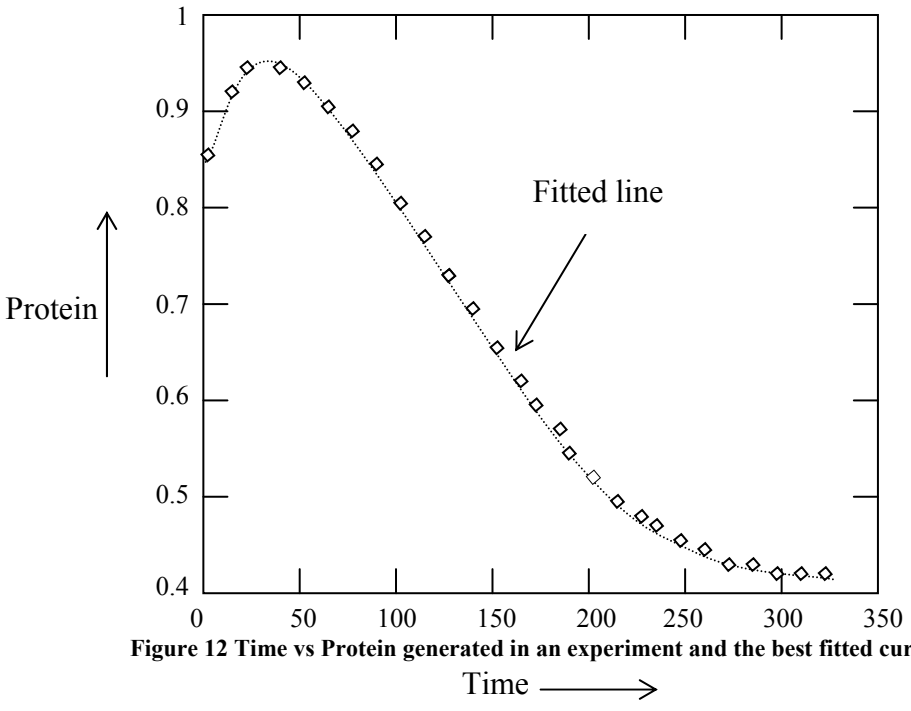


Figure 12 Time vs Protein generated in an experiment and the best fitted curve

in the Figure 12. The Figure 12 indicates that the fitted regression curve provides a very good relationship between the time and protein generated in that experiment. As before the prediction curve, that is, the curve (17) can be used easily for prediction purposes also. For example suppose we want to estimate the expected protein generation at the time 115 minutes after the experiment, then using (17), we obtain

$$\hat{y} = 0.375 + 1.936e^{-0.013\times 115} - 1.465e^{-0.013\times 115} = 0.698.$$

Therefore, at the 115 minutes the expected protein generation is 0.698 gms.

Some points we should remember about the non-linear regression model is that we have to know the functional form of the relationship, but the parameters involved are unknown. Usually the functional forms are obtained from the physical nature of the process and they are available. If they are completely unknown it is difficult to use the non-linear regression model. In that case we need to try with some guess models but they are not very easy and they are not pursued here. Another important issue is the choice of the guess value of the iterative process. This is also a difficult problem. Usually from the prior experimental results we may have to try some trial and error method to find the initial guess values.

Check your progress 4

E1 Data on the amount of heat generated by friction were obtained by Count Rumford in 1798. A bore was fitted into a stationery cylinder and pressed against the bottom by means of a screw. The bore was turned by a team of horses for 30 minutes and Rumford measured the temperature at small in intervals of time. They are provided in the Table below:

Table 8

Time (min)	Temperature ($°F$)	Time (min)	Temperature ($°F$)
4	126	24	115
5	125	28	114
7	123	31	113
12	120	34	112
14	119	37.5	111
16	118	41	110
20	116		

(1) Plot the time versus temperature and convince yourself that the linear regression model is not the correct model in this case.

(2) A model based on Newton's law of cooling was proposed as

$$f(t,\theta) = 60 + 70e^{-\theta t}$$

Using an initial guess of $\theta^{(0)} = 0.01$, find the least squares estimate of θ.
(3) Based on the fitted least squares regression line find the expected temperature at the time 15th minute after starting the experiment.

3.4 SUMMARY

In this unit you have seen :

- that regression analysis is an important technique, which can be used to verify the results of any experiment.
- How to determine the relationship between a dependent and an independent variable by using the Scatter diagram
- that by knowing the technique of regression you have an edge to analyse the results in an organized way. Further this analysis is smoothened by application of the concepts like least square estimation, goodness to fit and residual analysis.
- that many times the data obtained by conducting an experiment does not follow the linear relation. So, to handle such aspects we have also discussed the concept of non linear regression, under we have emphasized least square estimation technique.
- Formulas and applications of following topics:
    - Simple Linear Regression
        - Least Squares Estimation
        - Goodness to Fit
        - Residual Analysis
    - Non-Linear Regression
        - Least Squares Estimation

3.5 SOLUTIONS

Check you r progress 1

E 1:

(1) Since both the regression lines pass through the point $(\bar{x}, \bar{y})$, we have

$8\bar{x} - 10\ \bar{y} + 66 = 0$

$40\bar{x} - 18\ \bar{y} - 214 = 0$

Solving we get, $\bar{x} = 13$

$\bar{y} = 17.$

Let $8x - 10y + 66 = 0$ and $40x - 18y - 214 = 0$
Be the lines of regression of y and x and x on y respectively. Now, we put them in the following form.

$$= \frac{8}{10}x + \frac{66}{10} \text{ and } x = \frac{18}{40}y + \frac{214}{40} \qquad (4)$$

$\therefore$ byx = regression coeff of y on x = $\frac{8}{10} = \frac{4}{5}$

bxy = regression coeff of x on y = $\frac{18}{40} = \frac{9}{20}$

Hence, r^2 = bxy. byx = $\frac{4}{5} \quad \frac{9}{20} \quad = \quad \frac{9}{25}$

So r = $\pm \quad \frac{3}{5} \quad = \quad \pm \quad 0.6$

Since, both the regression coeff are +ve, we take r = + 0.6

(3) We have, byx = $r\frac{\sigma_y}{\sigma_x}$ $\Rightarrow$ $\frac{4}{5} \quad = \quad \frac{3}{5} \quad x \quad \frac{\sigma_y}{3}$

$\therefore \quad \sigma_y = 4$

Remarks (i) had we taken 8x – 10y + 66 = 0 as regression equation of x on y and 40x – 18y = 214, as regression equation of y on x.

Then bxy = $\frac{10}{8}$ and byx = $\frac{40}{18}$

or r^2 = bxy byx = $\frac{10}{8}$ x $\frac{40}{18} = 2.78$

so r = $\pm$ 1.66

Which is wrong as r lies between $\pm$ 1.

E 2

Observation	(x) Relative humidity, (%)	(y) Solvent Evaporation, (%) wt
1	35.3	11.0
2	29.7	11.1
3	30.8	12.5
4	58.8	8.4
5	61.4	9.3
6	71.3	8.7
7	74.4	6.4
8	76.7	8.5
9	70.7	7.8
10	57.5	9.1
11	46.4	8.2
12	28.9	12.2
13	28.1	11.9
14	39.1	9.6
15	46.8	10.9
16	48.5	9.6
17	59.3	10.1
18	70.0	8.1

19	70.0	6.8
20	74.4	8.9
21	72.1	7.7
22	58.1	8.5
23	44.6	8.9
24	33.4	10.4
25	28.6	11.1

Summary statistics for these data are

$n = 25$ $\quad \Sigma x = 1314.90$ $\quad \Sigma y = 235.70$
$\Sigma x^2 = 76.308.53$ $\quad \Sigma y^2 = 2286.07$ $\quad \Sigma xy = 11824.44$

To estimate the simple linear regression line, we estimate the slope β_1 and intercept β_0. these estimates are

$$=\beta_1 => \hat{\beta}_1 = b_1 = \frac{n\Sigma xy - [(\Sigma x)(\Sigma y)]}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{25(11,824.44) - [(1314.90)(235.70)]}{25(76,308.53) - (1314.90)^2}$$

$$= -.08$$

$$\beta_0 => \hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x}$$

$$= 9.43 - (-.08)\,(52.60) = 13.64$$

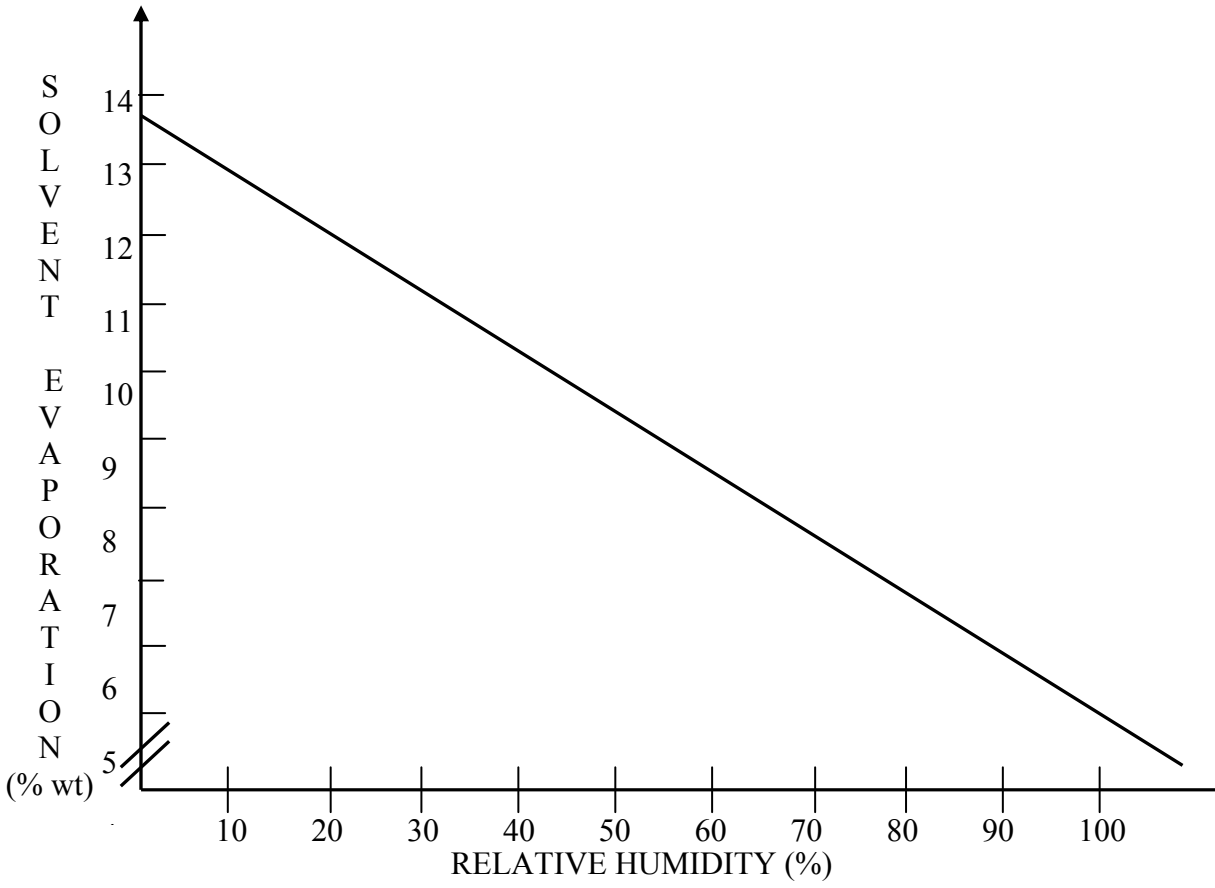


Figure: A graph of the estimated line of regression of Y, the extent of evaporation, on X, the relative humidity

Hence, the estimated regression equation is

$\mu_{Y|x} = \hat{y} = 13.64 - .08x$

The graph of this equation is shown in Figure above. To predict the extent of solvent evaporation when the relative humidity is 50 %, we substitute the value 50 for x in the equation.

$\hat{y} = 13.64 - .08x$

to obtain $\hat{y} = 13.64 - .08(50) = 9.64$. That is, when there relative humidity is 50 % we predict that 9.64% of the solvent, by weight, will be lost due to evaporation.

Recall from elementary calculus that the slope of a line gives the change in y for a unit change in x. If the slope is positive, then as x increases so does y; as x decreases, so does y. If the slope is negative, things operate in reverse. An increase in x signals a decrease in y, whereas a decrease in x yields an increase in y.

Check your progress 2

E1 $R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = (138.076) / [(2.6929 * 7839.93)^{1/2}] = 0.9504$

So, $R^2 = 0.9033$

Check your progress 3

E1 Residual plots are helpful in spotting potential problems. How ever theyare not always easy to interpret. Residual patterns are hard to spot with small data set except in extreme cases, residual plots are most useful with fairly large collection of data.

Check your progress 4

E1 Refer to example solved in the section 3.3.1

Table 1: The distribution function of standard normal random variable

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

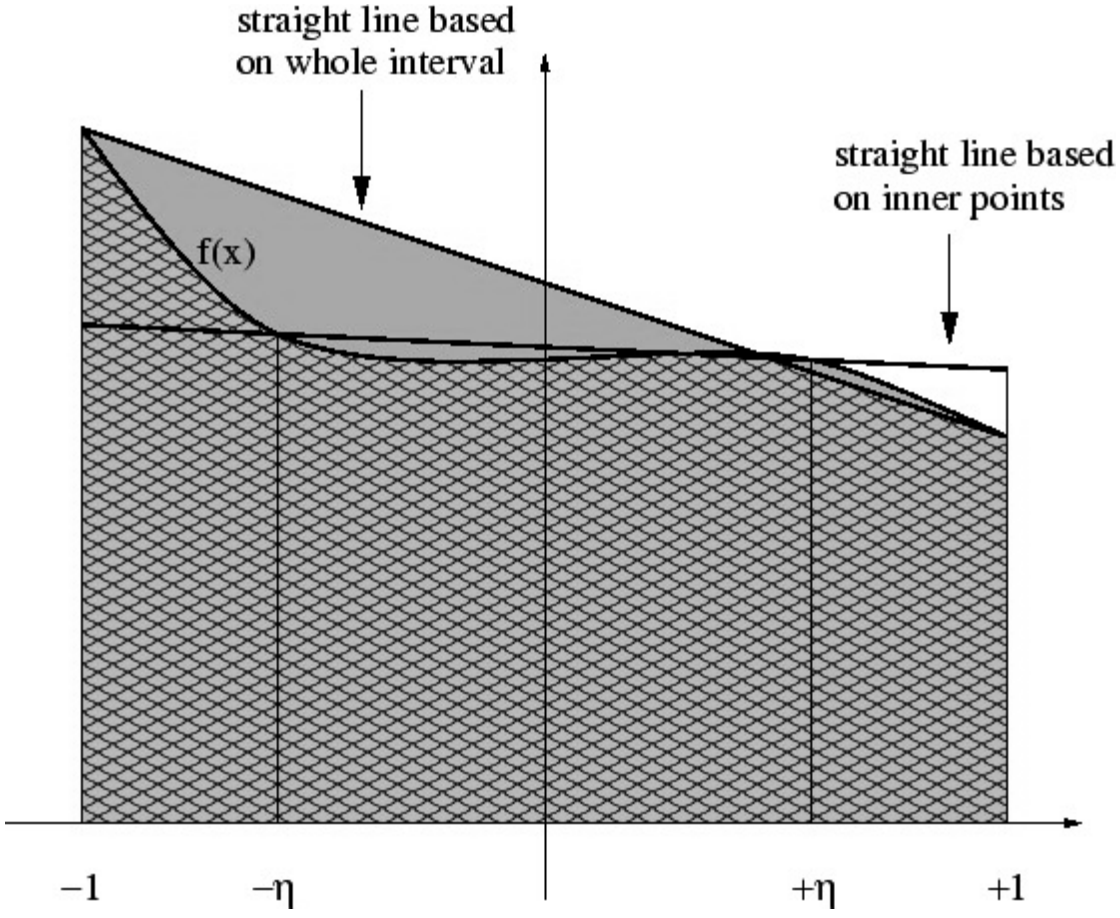
Table 2: The critical values of chi-square distribution. The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from one, and then look it up (i.e., 0.05 on the left is 0.95 on the right).

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	-	-	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.597	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.19	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.17	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.341	46.928

Gaussian Quadrature

The numerical integration methods described so far are based on a rather simple choice of evaluation points for the function $f(x)$. They are particularly suited for regularly tabulated data, such as one might measure in a laboratory, or obtain from computer software designed to produce tables. If one has the freedom to choose the points at which to evaluate $f(x)$, a careful choice can lead to much more accuracy in evaluating the integral in question. We shall see that this method, called Gaussian or Gauss-Legendre integration, has one significant further advantage in many situations. In the evaluation of an integral on the interval α to β, it is not necessary to evaluate $f(x)$ at the endpoints, ie. at α or β, of the interval. This will prove valuable when evaluating various *improper* integrals, such as those with infinite limits.

Figure 1.0: Comparing trapezoid rule integration and Gaussian integration.



We begin with a simple example illustrated in Figure.

The simplest form of Gaussian Integration is based on the use of an optimally chosen polynomial to approximate the integrand $f(t)$ over the interval [-1,+1]. The details of the determination of this polynomial, meaning determination of the coefficients of t in this

polynomial, are beyond the scope of this presentation. The simplest form uses a uniform weighting over the interval, and the particular points at which to evaluate $f(t)$ are the roots of a particular class of polynomials, the Legendre polynomials, over the interval. It can be shown that the best estimate of the integral is then:

$$\int_{-1}^{1} f(t)dt = \sum_{i=1}^{n} w_i f(t_i)$$

where t_i is a designated evaluation point, and w_i is the *weight* of that point in the sum. If the number of points at which the function $f(t)$ is evaluated is n, the resulting value of the integral is of the same accuracy as a simple polynomial method (such as Simpson's Rule) of about twice the degree (ie. of degree $2n$). Thus the carefully designed choice of function evaluation points in the Gauss-Legendre form results in the same accuracy for about half the number of function evaluations, and thus at about half the computing effort.

Gaussian quadrature formulae are evaluating using abscissae and weights from a table like that included here. The choice of value of n is not always clear, and experimentation is useful to see the influence of choosing a different number of points. When choosing to use n points, we call the method an ``n-point Gaussian'' method.

Gauss-Legendre Abscissae and Weights

n	Values of t	Weights	Degree
2	$\pm\sqrt{\frac{1}{3}}$	1.0	3
3	0.0	0.88888889	
	$\pm$ 0.77459667	0.55555555	5
4	$\pm$ 0.33998104	0.65214515	7
	$\pm$ 0.86113631	0.34785485	
5	0.0	0.56888889	9
	$\pm$ 0.53846931	0.47862867	
	$\pm$ 0.90617985	0.23692689	
6	$\pm$ 0.23861918	0.46791393	11

	$\pm$ 0.66120939	0.36076157	
	$\pm$ 0.93246951	0.17132449	
7	0.0	0.41795918	13
	$\pm$ 0.40584515	0.38183005	
	$\pm$ 0.74153119	0.27970539	
	$\pm$ 0.94910791	0.12948497	
8	$\pm$ 0.18343464	0.36268378	15
	$\pm$ 0.52553241	0.31370665	
	$\pm$ 0.79666648	0.22238103	
	$\pm$ 0.96028986	0.10122854	
10	$\pm$ 0.14887434	0.29552422	19
	$\pm$ 0.43339539	0.26926672	
	$\pm$ 0.67940957	0.21908636	
	$\pm$ 0.86506337	0.14945135	
	$\pm$ 0.97390653	0.06667134	

The Gauss-Legendre integration formula given here evaluates an estimate of the required integral on the interval for t of [-1,+1]. In most cases we will want to evaluate the integral on a more general interval, say $[\alpha, \beta]$. We will use the variable x on this more general interval, and linearly map the $[\alpha, \beta]$ interval for x onto the [-1,+1] interval for t using the linear tranformation:

$$x = c + mt \quad \text{where} \quad c = \frac{1}{2}(b+a) \quad \text{and} \quad m = \frac{1}{2}(b-a)$$

It is easily verified that substituting t = -1 gives $x = a$ and t = 1 gives $x = b$. We can now write the integral as:

$$I = \int_a^b f(x)\,dx = m\int_{-1}^{+1} f(c+mt)\,dt$$

The factor of *m* in the second integral arises from the change of the variable of integration from *x* to *t*, which introduces the factor $\frac{dx}{dt}$. Finally, we can write the Gauss-Legendre estimate of the integral as:

$$I = \int_a^b f(x)\,dx = m\sum_{i=1}^{n} w_i f(c+mt_i)$$

Consider the evaluation of the integral:

$$I = \int_0^{\pi/2} \sin x\,dx$$

whose value is 1, as we can obtain by explicit integration. Applying the 2-point Gaussian method, and noting that both *c* and *m* are $\frac{\pi}{4}$, the table allows us to calculate an approximate value for the integral. The result is 0.998473, which is pretty close to the exact value of one. The calculation is simply:

$$I \approx \frac{\pi}{4}\left(1.0\times\sin\left(\left(1-\frac{1}{\sqrt{3}}\right)\frac{\pi}{4}\right)+1.0\times\sin\left(\left(1+\frac{1}{\sqrt{3}}\right)\frac{\pi}{4}\right)\right)$$

While this example is quite simple, the following table of values obtained for *n* ranging from 2 to 10 indicates how accurate the estimate of the integral is for only a few function evaluations. The table includes a column of values obtained from Simpson's $\frac{1}{3}$ rule for the same number of function evaluations. The Gauss-Legendre result is correct to almost twice the number of digits as compared to the Simpson's rule result for the same number of function evaluations.

N	Gauss-Legendre	Simpson's 1/3
2	0.9984726135	1.0022798775
4	0.9999999770	1.0001345845
6	0.9999999904	1.0000263122
8	1.0000000001	1.0000082955
10	0.9999999902	1.0000033922

Example: Evaluate the integral

$$I = \int \frac{dx}{1+x}$$

Using Gauss-Legendre three point formula.

First we transform the interval [0, 1] to the interval [-1, 1], Let t=ax+b.

We have

$$I = \int \frac{dx}{1+x}$$

$$-1 = b, 1 = a + b$$

or

$$a = 2, b = -1,$$

$$t = 2x - 1.$$

$$I = \int_0^1 \frac{dx}{1+x} = \int_{-1}^1 \frac{dt}{t+3}.$$

$$I = \frac{1}{9}[8(1/0+3) + 5(1/3+\sqrt{3/5}) + 5(1/3-\sqrt{3/5})]$$

$$= \frac{131}{189} = 0.693122$$

The exact solution is

$$I = In2 = 0.693147$$

LU DECOMPOSITION METHOD

The Gauss Elimination Method has the disadvantage that all right-hand sides (i.e. all the ***b*** *vectors of interest for a given problem) must be known in advance for the elimination step to proceed. The LU Decomposition Method outlined here has the property that the matrix modification (or decomposition) step can be performed independent of the right hand side vector. This feature is quite useful in practice - therefore, the LU Decomposition Method is usually the Direct Scheme of choice in most applications.*

To develop the basic method, let's break the coefficient matrix into a product of two matrices,

$$\boldsymbol{A = L\,U} \qquad (3.12)$$

where ***L*** *is a lower triangular matrix and* ***U*** *is an upper triangular matrix.*

Now, the original system of equations,

$$\boldsymbol{A\,x = b} \qquad (3.13)$$

becomes

$$\boldsymbol{LU\,x = b} \qquad (3.14)$$

This expression can be broken into two problems,

$$\boldsymbol{Ly = b} \quad \textit{and} \quad \boldsymbol{U\,x = b} \qquad (3.15)$$

The rationale behind this approach is that the two systems given in eqn. (3.15) are both easy to solve; one by forward substitution and the other by back substitution. In particular, because ***L*** *is a lower diagonal matrix, the expression,* $\boldsymbol{Ly = b}$ *, can be solved with a simple forward substitution step. Similarly, since* ***U*** *has upper triangular form,* $\boldsymbol{U\,x = b}$ *can be evaluated with a simple back substitution algorithm.*

Thus the key to this method is the ability to find two matrices, ***L*** *and* ***U****, that satisfy eqn. (3.12). Doing this is referred to as the Decomposition Step and there are a variety of algorithms available. Three specific approaches are as follows:*

Doolittle Decomposition*:*

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ \ell_{21} & 1 & 0 & 0 \\ \ell_{31} & \ell_{32} & 1 & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \qquad (3.16)$$

Because of the specific structure of the matrices, a systematic set of formulae for the components of ***L*** *and* ***U*** *results.*

Crout Decomposition*:*

$$\begin{bmatrix} \ell_{11} & 0 & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & \ell_{44} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \quad (3.17)$$

The evaluation of the components of ***L*** *and* ***U*** *is done in a similar fashion as above.*

Cholesky Factorization*:*

For symmetric, positive definite matrices, where

$$\boldsymbol{A = A^T \quad and \quad x^T A\, x > 0 \quad for\ x \neq 0} \qquad (3.18)$$

then,

$$\boldsymbol{U = L^T} \quad and \quad \boldsymbol{A = L\, L^T} \qquad (3.19)$$

and a simple set of expressions for the elements of ***L*** *can be obtained (as above).*

Once the elements of ***L and U*** *are available (usually stored in a single NxN matrix), the solution step for the unknown vector* ***x*** *is a simple process [as outlined above in eqn. (3.15)].*

A procedure for decomposing an $N \times N$ matrix **A** into a product of a ower triangular matrix L and an upper triangular matrix U,

$$L\,U = A.$$

Written explicitly for a 3×3 matrix the decomposition is

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\begin{bmatrix} l_{11}\,u_{11} & l_{11}\,u_{12} & l_{11}\,u_{13} \\ l_{21}\,u_{11} & l_{21}\,u_{12} + l_{22}\,u_{22} & l_{21}\,u_{13} + l_{22}\,u_{23} \\ l_{31}\,u_{11} & l_{31}\,u_{12} + l_{32}\,u_{22} & l_{31}\,u_{13} + l_{32}\,u_{23} + l_{33}\,u_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

(3.20)

This gives three types of equations

$$i < j \quad l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{ii}u_{ij} = a_{ij} \tag{3.21}$$
$$i = j \quad l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{ii}u_{jj} = a_{ij} \tag{3.22}$$
$$i > j \quad l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{ij}u_{jj} = a_{ij}. \tag{3.23}$$

This gives N^2 equations for $N^2 + N$ unknown (the decomposition is not unique), and can be solved using either using Doolittle or Crout's method.

Doolittle Method : Here $l_{ii} = 1$, i = 1 to N. In this case , equation (3.20) gives

$$u_{1j} = a_{1j} \qquad j = 1 \text{ to } N$$

$$l_{i1} = a_{i1} / a_{11}, \qquad i = 2 \text{ to } N$$

$$u_{2j} = a_{2j} - l_{21} . u_{1j}, \quad j=2 \text{ to } N$$

$$l_{i2} = (a_{i2} - l_{i1} u_{12}) / u_{22}, \quad i = 3 \text{ to } N,$$

and so on

Crout's Method : Here $u_{ii} = 1$, i = 1 to N . In this case , we get

$$l_{i1} = a_{i1}, \qquad i=1 \text{ to } N$$

$$u_{1j} = a_{1j} / a_{11}, \qquad j= 2 \text{ to } N$$

$$l_{i2} = a_{i2} - l_{i1} u_{12}, \qquad i = 2 \text{ to } N$$

$$u_{2j} = (a_{2j} - l_{21} u_{1j}) / l_{22}, \quad j= 3 \text{ to } N,$$

and so on

Example 7: *Given the following system of linear equations, determine the value of each of the variables using the LU decomposition method.*

$$\begin{aligned} 6x_1 - 2x_2 &= 14 \\ 9x_1 - x_2 + x_3 &= 21 \\ 3x_1 - 7x_2 + 5x_3 &= 9 \end{aligned} \tag{3.24}$$

Solution *:*

Upper Triangular	***Explanation of Step***	***Lower Triangular***
$\begin{bmatrix} 6 & -2 & 0 \\ 9 & -1 & 1 \\ 3 & 7 & 5 \end{bmatrix}$	*<--- Beginning Matrix* *Matrix Storing Elementary Row Operations --->*	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$\begin{bmatrix} 1 & -1/3 & 0 \\ 9 & -1 & 1 \\ 3 & 7 & 5 \end{bmatrix}$	*In order to force a value of 1 at position (1,1), we must multiply row 1 by 1/6. Thus storing its reciprocal, 6, in position (1,1) in the lower matrix.*	$\begin{bmatrix} 6 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$\begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 2 & 1 \\ 0 & 8 & 5 \end{bmatrix}$	*Introducing zeros to positions (2,1) and (3,1) require multiplications by -9 and -3 respectively. So we will store the opposite of these numbers in their respective locations.*	$\begin{bmatrix} 6 & 0 & 0 \\ 9 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix}$
$\begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 1/2 \\ 0 & 8 & 5 \end{bmatrix}$	*On to the next position in the main diagonal, (2,2). To replace the value in this position with a 1, multiply row 2 by 1/2, thus storing a 2 (the reciprocal) in position (2,2) in the lower triangular matrix.*	$\begin{bmatrix} 6 & 0 & 0 \\ 9 & 2 & 0 \\ 3 & 0 & 0 \end{bmatrix}$
$\begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$	*Replacing the position under the leading 1, position (3,2), with a zero can be done with a multiplication of -8. We will then store 8, the opposite of -8, in the lower matrix at that position.*	$\begin{bmatrix} 6 & 0 & 0 \\ 9 & 2 & 0 \\ 3 & -8 & 0 \end{bmatrix}$
$\begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$	*Only a multiplication of 1 is necessary to introduce a 1 to the next diagonal position. In fact nothing is being done to the upper triangular matrix, but we need the 1 in the lower matrix to show that.*	$\begin{bmatrix} 6 & 0 & 0 \\ 9 & 2 & 0 \\ 3 & -8 & 1 \end{bmatrix}$

If a matrix A can be decomposed into an LU representation, then A is equal to the product of the lower and upper triangular matrices. This can be shown with one matrix multiplication.

$$\begin{bmatrix} 6 & -2 & 0 \\ 9 & -1 & 1 \\ 3 & 7 & 5 \end{bmatrix} = \begin{bmatrix} 6 & 0 & 0 \\ 9 & 2 & 0 \\ 3 & -8 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix} \qquad (3.25)$$

Solving Systems of Equations using the LU decomposition.
Systems of linear equations can be represented in a number of ways. In the Gauss-Jordan elimination method, the system was represented as an augmented matrix. In this method, we will represent the system as a matrix equation.

1. *Rewrite the system $A\boldsymbol{x} = \boldsymbol{b}$ using the LU representation for A. Making the system $LU\boldsymbol{x} = \boldsymbol{b}$.*

$$\begin{bmatrix} 6 & 0 & 0 \\ 9 & 2 & 0 \\ 3 & -8 & 1 \end{bmatrix}\begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 14 \\ 21 \\ 9 \end{bmatrix}$$

2. *Define a new column matrix **y** so that U**x** = **y**.*

$$\begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$$

3. *Rewrite step one with the substitution from step two yielding L**y** = **b**.*

$$\begin{bmatrix} 6 & 0 & 0 \\ 9 & 2 & 0 \\ 3 & -8 & 1 \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 14 \\ 21 \\ 9 \end{bmatrix}$$

4. *Solve step three for **y** using forward substitution.*

   $y_1 = 7/3, \quad y_2 = 29/6, \quad y_3 = 33/2$

5. *Using the results from step four, solve for **x** in step two using back substitution.*

$$\begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 2 \end{bmatrix}$$

   $x_1=43/36, x_2=-41/12, x_3=33/2$

METHOD OF SUCCESSIVE ITERATION

The first step in this method is to write the equation in the form

$$x = g(x) \qquad (14)$$

For example, consider the equation $x^2 - 4x + 2 = 0$ *. We can write it as*

$$x = \sqrt{4x - 2} \qquad (15)$$

$$x = (x^2 + 2)/4 \qquad (16)$$

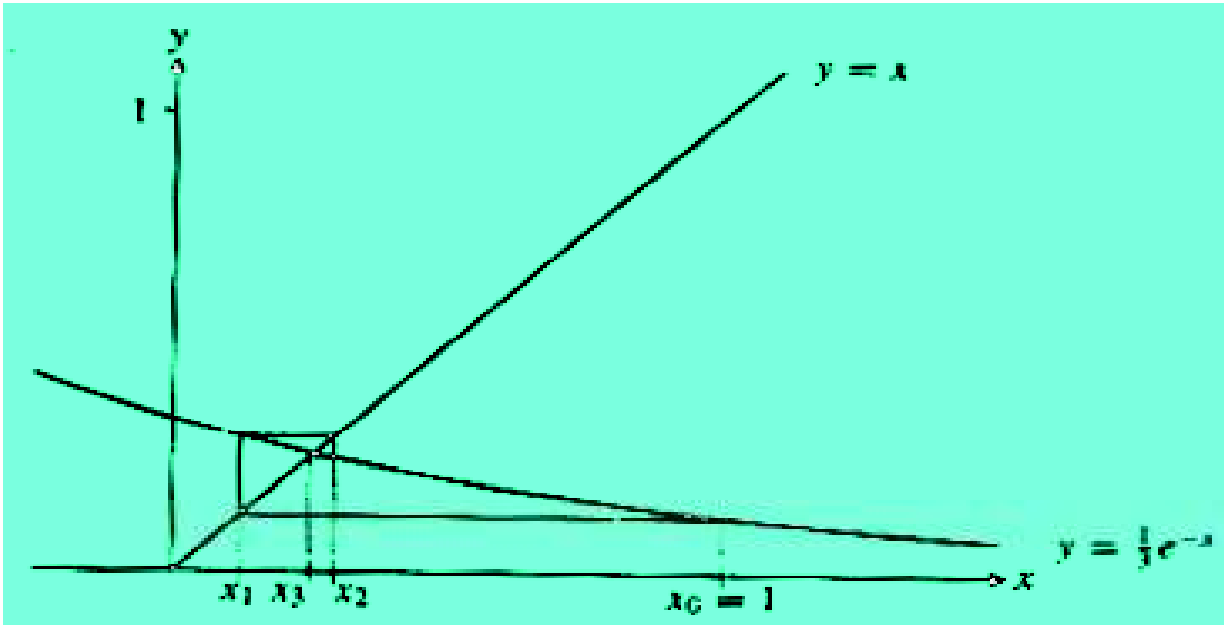
$$x = \frac{2}{4 - x} \qquad (17)$$

Thus, we can choose form (1) in several ways. Since $f(x) = 0$ *is the same as* $x = g(x)$ *, finding a root of* $f(x) = 0$ *is the same as finding a root of* $x = g(x)$*, i.e. finding a fixed point* α *of* $g(x)$ *such that* $\alpha = g(\alpha)$ *. The function* $g(x)$ *is called an* **iterative function** *for solving* $f(x) = 0$ *.*

If an initial approximation x_0 *to a root* α *is provided , a sequence* x_1 *,* x_2 *,..... may be defined by the iteration scheme*

$$x_{n+1} = g(x_n) \qquad (18)$$

with the hope that the sequence will converge to α *. The successive iterations are interpreted graphically , shown in the following figure*



Convergence will certainly occur if , for some constant M such that 0< M <1, the inequality

$$| g(x) - g(\alpha) | \leq M | x - \alpha | \qquad (19)$$

holds true whenever $|x- \alpha | \leq | x_0 - \alpha |$ *. For, if (6) holds, we find that*

$$| x_{n+1} - \alpha | = | g(x_n) - \alpha | = | g(x_n) - g(\alpha) | \leq M | x_n - \alpha | \qquad (20)$$

Proceeding further,

$$| x_{n+1} - \alpha | \leq M | x_n - \alpha | \leq M^2 | x_{n-1} - \alpha | \leq M^3 | x_{n-2} - \alpha | \qquad (21)$$

Continuing in this manner, we conclude that

$$| x_n - \alpha | \leq M^n | x_0 - \alpha | \qquad (22)$$

Thus, $\lim x_n = \alpha$ *, as* $\lim M^n = 0$ *.*

Condition (6) is clearly satisfied if function g(x) possesses a derivative g'(x) such that $| g'(x) | < 1$ *for* $| x-\alpha | < | x_0 - \alpha |$ *.*

If x_n *is closed to* α *, then we have*

$$\begin{aligned} | x_{n+1} - \alpha | &= | g(x_n) - g(\alpha) | \\ &\leq g'(\xi) | x_n - \alpha | \end{aligned} \qquad (23)$$

for some ξ *between* x_0 *and* α *.*

Therefore, condition for convergence is $| g'(\xi) | < 1$.

Example 9 :

Lets consider $f(x) = x^3 + x - 2$, which we can see has a single root at x=1. There are several ways $f(x)$=0 can be written in the desired form, x=$g(x)$.

The simplest is

$$x_{n+1} = x_n + f(x_n) = x_n^3 + 2x_n - 2$$

In this case, $g'(x) = 3x^2 + 2$, and the convergence condition is

$$1 > |g'(x)| = 3x^2 + 2, \qquad -1 > 3x^2$$

Since this is never true, this doesn't converge to the root.

An alternate rearrangement is

$$x_{n+1} = 2 - x_n^3$$

This converges when

$$1 > |g'(x)| = |-3x^2|, \qquad x^2 < \frac{1}{3}, \qquad |x| < \frac{1}{\sqrt{3}}$$

Since this range does not include the root, this method won't converge either.

Another obvious rearrangement is

$$x_{n+1} = \sqrt[3]{2 - x_n}$$

In this case the convergence condition becomes

$$\frac{1}{3}\left|(2 - x_n)^{-\frac{2}{3}}\right| < 1, \qquad (2 - x_n)^{-2} < 3^3, \qquad |x_n - 2| > \sqrt{27}$$

Again, this region excludes the root.

Another possibility is obtained by dividing by x^2+1

$$x_{n+1} = \frac{2}{x_n^2 + 1}$$

In this case the convergence condition becomes

$$\frac{4|x|}{(1 + x^2)^2} < 1, \qquad 4|x| < (1 + x^2)^2$$

Consideration of this ineqaulity shows it is satisified if $x>1$, so if we start with such an x, this will converge to the root.

Clearly, finding a method of this type which converges is not always straightforwards

The Successive Overrelaxation Method

The Successive Over relaxation Method, or SOR, is devised by applying extrapolation to the Gauss-Seidel method. This extrapolation takes the form of a weighted average between the previous iterate and the computed Gauss-Seidel iterate successively for each component:

$$x_i^{(k)} = \omega \bar{x}_i^{(k)} + (1-\omega)x_i^{(k-1)} \qquad (3.38)$$

(where $\bar{x}$ denotes a Gauss-Seidel iterate, and ω is the extrapolation factor). The idea is to choose a value for ω that will accelerate the rate of convergence of the iterates to the solution.

In matrix terms, the SOR algorithm can be written as follows:

$$x^{(k)} = (D-\omega L)^{-1}(\omega U + (1-\omega)D)x^{(k-1)} + \omega(D-\omega L)^{-1}b. \qquad (3.39)$$

Example 13 : Solve the 3 by 3 system of linear equations **Ax = b** where

$$A = \begin{bmatrix} 4 & -2 & 0 \\ -2 & 6 & -5 \\ 0 & -5 & 11 \end{bmatrix} \qquad b = \begin{bmatrix} 8 \\ -29 \\ 43 \end{bmatrix}$$

by SOR method .

Solution : For SOR iterations , the system can be written as

$$x_1^{(new)} = (1-\omega)x_1^{(old)} + \omega(\frac{1}{2}x_2^{(old)} + 2)$$

$$x_2^{(new)} = (1-\omega)x_2^{(old)} + \omega(\frac{1}{3}x_1^{(new)} + \frac{5}{6}x_3^{(old)} - \frac{29}{6})$$

$$x_3^{(new)} = (1-\omega)x_3^{(old)} + \omega(\frac{5}{11}x_1^{(new)} + \frac{43}{11})$$

Start with $x^0 = (0, 0, 0)^T$, for $\omega = 1.2$ we get the following solution

```
>> SOR_f(A,b,x0,1.2,0.001,50)
   1.0000   2.4000  -4.8400   2.0509

   2.0000  -0.9840  -3.1747   2.5491

   3.0000   0.6920  -2.3392   2.9052

   4.0000   0.8581  -2.0838   2.9733

   5.0000   0.9781  -2.0187   2.9951

   6.0000   0.9931  -2.0039   2.9989

   7.0000   0.9991  -2.0007   2.9998

SOR method converged
   8.0000   0.9997  -2.0001   3.0000
```

Infact the required number of iterations for different values of relaxation parameter ω for tolerence value 0 .00001 is as follows

Ω	0.8	0.9	1.0	1.2	1.25	1.3	1.4
No. of iterations	44	36	29	18	15	13	16